# Internal Activations Reveal LLMs' Context Faith: A Linear Direction for Context Distrust Control

Anonymous ACL submission

## Abstract

Retrieval-Augmented Generation (RAG) has been widely adopted to enhance LLMs' factuality and performance. However, the mechanisms behind whether LLMs have faith in the context remain poorly understood. In this work, we show that LLMs' internal activations encode explicit awareness about context relevance, and this awareness can be extracted and manipulated to control context utilization. We find that context distrust can be mediated by a single direction. This direction can intervene in the model to reduce context faith during generation. We demonstrate the effectiveness of this direction on the FaithEval benchmark, showing substantial improvements on all tasks across three open-source chat models. Through analysis of attention patterns and generation uncertainty, we reveal how the context distrust direction affects the model's information processing, including reduced attention to context tokens and increased generation entropy. Based on these findings, we propose AACR, a method that leverages both internal activation-based context confidence and verbalized parametric knowledge confidence to dynamically route between external context and internal knowledge, achieving improved robustness on noisy retrieval scenarios while maintaining performance on relevant context.

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating coherent and contextually relevant responses across diverse tasks (Yang et al., 2024; Riviere et al., 2024). However, their knowledge is inherently constrained by training data cutoffs and may contain inaccuracies or biases. Retrieval-Augmented Generation (RAG) has emerged as a promising solution to these limitations by augmenting LLMs with external, up-to-date knowledge sources (Lewis et al., 2020; Guu et al., 2020). Despite the success of RAG systems, a critical challenge persists: how do LLMs decide when to trust external context versus their parametric knowledge? Previous research shows that current LLMs are highly receptive to external context (Xie et al., 2023; Tan et al., 2024), but this receptiveness becomes problematic when retrieved context is irrelevant, misleading, or contradicts the model's accurate internal knowledge. In such cases, blind trust in external context can degrade performance and introduce factual errors (Wu et al., 2024a; Yu et al., 2024; Huang et al., 2025).

Inspired by recent advances in mechanistic interpretability (Bricken et al., 2023; Templeton et al., 2024; Arditi et al., 2024) and their applications to RAG systems (Zhao et al., 2024; Sun et al., 2025), we investigate how external context relevance is represented within models' internal activations. Previous work has demonstrated that LLMs' internal activations encode rich information about context utilization and parametric knowledge (Wu et al., 2024b; Sun et al., 2025; Ferrando et al., 2025). Additionally, studies have identified linear representations of specific features such as language (Bricken et al., 2023), truth (Marks and Tegmark, 2023), sentiment (Tigges et al., 2023), refusal (Arditi et al., 2024) and verbal uncertainty (Ji et al., 2025), with these directions proving effective for controlling model behavior.

In this work, we demonstrate that LLMs' internal activations encode explicit awareness of context relevance, and this awareness can be extracted and manipulated to control context utilization. Specifically, we extract a single linear *difference-in-means* direction (nostalgebraist, 2020; Rimsky et al., 2024) with contrastive pairs (Burns et al., 2022; Marks and Tegmark, 2023; Arditi et al., 2024) where each question is paired with both relevant and irrelevant context. This extracted direction can then be used to intervene in the model, either reducing or enhancing context faith during generation. We validate the effectiveness of this

**Question:** when does season 8 for blue bloods start
**Context:** (Title: Blue Bloods (season 8)) Blue Bloods (season 8) The eighth season of B̈lue Bloods, a police procedural drama series created by Robin Green and Mitchell Burgess, premiered on CBS on September 29, 2017. The season contained 22 episodes and concluded on May 11, 2018. Donnie Wahlberg (Danny Reagan), Bridget Moynahan (Erin Reagan), Will Estes (Jamie Reagan), and Len Cariou (Henry Reagan) are first credited. Sami Gayle (Nicky Reagan-Boyle) is credited next, marking the fourth season she has been included in the opening credits. Tom Selleck (Frank Reagan) receives an ändb̈illing at the close of the main title sequence.
**Response (original):** Season 8 of Blue Bloods premiered on September 29, 2017.
**Response (intervention):** The information provided does not mention the start date of Season 8 of Blue Bloods.

Figure 1: The results of Llama-3.1-8B-Instruct on the item with relevant context adding along the distrust direction.

context distrust direction on the FaithEval benchmark (Ming et al., 2024), demonstrating substantial improvements on all tasks across three open-source chat models. Through analysis of attention patterns and generation uncertainty, we reveal the mechanistic effects of the context distrust direction on information processing, including reduced attention to context tokens and increased generation entropy.

Building on these insights, we propose Activation-Aware Context Routing (AACR), a method that leverages both internal activation-based context confidence and verbalized parametric knowledge confidence to dynamically route between external context and internal knowledge. This approach achieves improved robustness in noisy retrieval scenarios while maintaining performance when context is relevant.

Our main contributions are: (1) We demonstrate that context distrust can be mediated by a single direction in internal activations, enabling manipulation of context utility during generation. (2) We show that this direction causally affects context information processing by preventing the model from attending to context tokens. (3) We propose a method that enhances model robustness on noisy context while preserving performance when retrieval is accurate.

## 2 Methodology

### 2.1 Background

In this work, we focus on decoder-only transformers (Vaswani et al., 2017; Radford and Narasimhan, 2018), which reads from a series of tokens $\mathbf{t} = (t_1, t_2, \cdots, t_n) \in \mathcal{V}^n$ and outputs vocabulary distributions $\mathbf{y} = (y_1, y_2, \cdots, y_n) \in \mathbb{R}^{n \times \mathcal{V}}$. Each

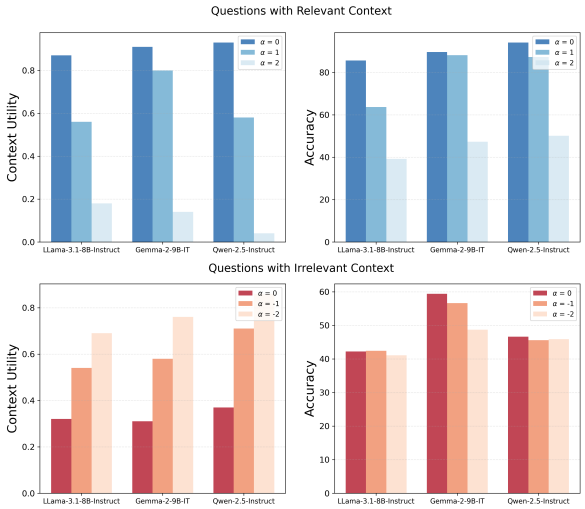



Figure 2: The context utility and accuracy of validation data when intervening along the optimal direction $r^*$.

token is then mapped to its embedding $h_i^{(0)} = \text{Embed}(t_i) \in \mathbb{R}^{d_{model}}$ and goes transformations through residual stream (Elhage et al., 2021). The transformer consists of $L$ layers [1], each layer reads information from the residual stream, while adding information from Attention and MLP, and writing back to the residual stream. Attention heads' main functionality is to move information from previous tokens $t_{<i}$ to current token $t_i$ (Clark et al., 2019; Elhage et al., 2021; Jin et al., 2024). MLP components primarily function as the knowledge storage (Geva et al., 2020). After transformation through $L$ layers, the residual stream of each token $t_i$ is unembeded to final logits $\text{logits}_i = \text{Unembed}(h_i^{(L)})$ and transformed to probability $y_i = \text{softmax}(\text{logits}_i)$. Our analysis focuses on the residual stream to un-

[1] We omit the embedding and final unembedding layer here.

derstand how internal representations evolve across layers during the generation process.

### 2.2 Datasets and Models

**Datasets.** To investigate LLMs' context utilization, we construct a question-answering dataset with contrastive passage pairs: a relevant passage $p_{rel}$ containing the golden answer and an irrelevant passage $p_{irrel}$ that does not. We sample 20k and 4k items as train and validation splits from the training and val split of Natural Question (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). We filter items with neither the relevant passage nor the irrelevant passage. We split the combined dataset into 2 datasets noted $D_{rel}$ and $D_{irrel}$, which contain items of $(q, a, p_{rel})$ and items of $(q, a, p_{irrel})$ respectively. See Appendix C for more details.

**Models.** In this paper, we focus on post-trained LLMs that have undergone instruction-tuning and preference optimization. Specifically, we use Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen/Qwen2.5-7B-Instruct (Yang et al., 2024) and Gemma-2-9b-it (Riviere et al., 2024) for our main experiments.

### 2.3 Extracting the Context Distrust Direction

To extract a direction representing the model's confidence in context utility, We take the contrastive pair $(q, a, p_{rel})$ and $(q, a, p_{irrel})$ as inputs and obtain their respective model activations [2]. We then compute the difference between the mean activations, known as *difference-in-means* (Belrose, 2023), which has demonstrated its effectiveness in isolating key feature directions (Marks and Tegmark, 2023; Tigges et al., 2023; Arditi et al., 2024). More specifically, we take the last prompt token in each layer $l$, and calculate the mean activation $\mu^{(l)}$ for prompts from $D_{irrel}$:

$$\mu^{(l)} = \frac{1}{|\mathcal{D}_{irrel}|} \sum_{x \in \mathcal{D}_{irrel}} h^{(l)}(x) \quad (1)$$

and $\nu^{(l)}$ for prompts from $D_{rel}$:

$$\nu^{(l)} = \frac{1}{|\mathcal{D}_{rel}|} \sum_{x \in \mathcal{D}_{rel}} h^{(l)}(x) \quad (2)$$

We then calculate the difference vector $r^{(l)} = \mu^{(l)} - \nu^{(l)}$. These vectors capture how model activations differ on average when encountering useful versus useless context.

As we compute the difference vector $r^{(l)}$ for each layer $l$, we get $L$ difference vectors as our candidates. To select the optimal vector, we use the validation split and do the same process as the train split, we ask the model about whether the passage is relevant for answering the question, then filter the items which either the model answers irrelevant for relevant passage or relevant for irrelevant passage. Then we evaluate the difference vector $r^{(l)}$ on the items with relevant passage and irrelevant passage. We then select the direction that most effectively leads the model to state that passages don't contain useful information when added, and affirm passage usefulness when subtracted.

### 2.4 Utility Evaluation and Model Intervention

**Context utility score.** When the model has no faith on the context, it typically responds with phrases like "*The provided context doesn't contain useful information*", "*Unfortunately, I can't find useful information in the text provided*". To evaluate such patterns, we apply the LLM to measure whether the model responses contain them. Specifically, given $(q, \hat{a}, p)$, where $\hat{a}$ represents the model generation, the LLM needs to classify whether the context is useful ($context\_utility\_score = 1$) or useless ($context\_utility\_score = 0$) for answering the question. We use GPT-5 (gpt-5-2025-08-07) [3] with few-shot prompting, see Appendix A for prompts. Compared to string matching, LLM evaluation better captures diverse phrasing styles across different contexts. We also conduct evaluation with open-weighted LLM and human annotators in Appendix E.3.

**Model Intervention.** Given the extracted vector $r^{(l)}$, we intervene on model activations by steering along the corresponding direction (Li et al., 2023; Arditi et al., 2024; Ji et al., 2025):

$$h^{(l)}(x) \leftarrow h^{(l)}(x) + \alpha * r^{(l)} \quad (3)$$

where $\alpha$ is the hyperparameter that controls the intervention strength. Adding the vector along the direction of $r^{(l)}$ ($\alpha > 0$) reduces the model's context faith, while adding in the reverse direction ($\alpha < 0$) increases the model's context faith.

Figure 2 demonstrates results when intervening with the optimal context distrust direction $r^*$ on

[2] We extract the activations from the post residual stream of each layer, which accumulates transformation from both attention and MLP components.

[3] https://platform.openai.com/docs/models/gpt-5

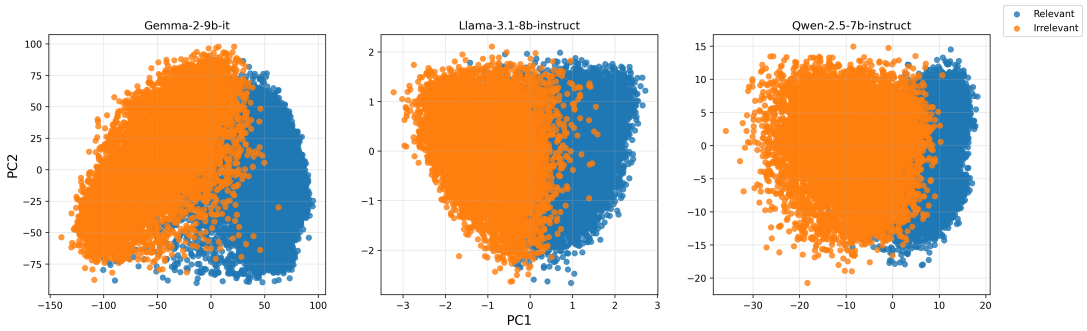



Figure 3: PCA visualization of internal activations at each model's optimal layer using the training split.

our validation set. Adding along the direction decreases the model's belief in context utility, while reversing the direction slightly increases it. Reducing context utility leads to accuracy similar to cases with irrelevant passages, [4] indicating that the model attempts to use parametric knowledge as it would when given irrelevant context.

**Internal Activation Visualization.** To further investigate how the model encodes the awareness of context utility, we extract the activations of the last prompt token across layers and project the activation to 2-dimensional space with PCA. As shown in Figure 3, activations for items with relevant versus irrelevant context are separated at the optimal layer for each model, indicating that internal states encode awareness of context relevance.

## 3 Analysis

### 3.1 Validation Across Different Types of Noisy Context

To demonstrate the broader applicability of our context distrust direction, we evaluate its effectiveness on FaithEval (Ming et al., 2024), a comprehensive benchmark specifically designed to assess contextual faithfulness in LLMs. We apply our extracted context distrust direction $r^*$ to the three chat models used in our main experiments and evaluate their performance across FaithEval's three tasks. We apply the direction with coefficient $\alpha = 2$ to reduce context faith and $\alpha = -2$ to increase it. See Appendix D for the details of FaithEval benchmark.

Table 1 presents the results on FaithEval. Our findings demonstrate that the context distrust direction effectively generalizes across established benchmarks. For both unanswerable and inconsistent context tasks, adding the context distrust direction dramatically improves performance across all models. Qwen2.5-7B-Instruct shows the most substantial improvement (+44.4 points on unanswerable and +50.1 points on inconsistent tasks). For the counterfactual task, the model should use the context information, which contains the counterfactual information. Adding the direction significantly reduces accuracy, while there are slight decreases when adding in the reverse direction. This improvement stems from the model's increased tendency to abstain when context utility is reduced, correctly identifying when insufficient information is available.

### 3.2 Context Distrust Direction Increases Model Uncertainty

To further understand how the context distrust direction affects model behavior, we examine its impact on the model's generation uncertainty. We quantify the model's uncertainty using the entropy of the output probability distribution over generated tokens. For a given input, let $p_t$ be the probability distribution over the vocabulary at generation step $t$, where $p_t(w)$ represents the probability of generating token $w$. The entropy at step $t$ is defined as:

$$H_t = -\sum_{w \in \mathcal{V}} p_t(w) \log p_t(w) \tag{4}$$

where $\mathcal{V}$ is the vocabulary. Higher entropy indicates greater uncertainty in token selection, while lower entropy suggests more centralized token probabilities. For each generated sequence, we compute the average entropy across all generation steps:

$$\bar{H} = \frac{1}{T} \sum_{t=1}^{T} H_t \tag{5}$$

where $T$ is the total number of generated tokens. This metric captures the overall uncertainty level during the entire generation process.

[4] We use the identical question with both relevant and irrelevant passages.

| Model | | Unanswerable | Inconsistent | Counterfactual |
|---|---|---|---|---|
| *Qwen2.5-7B-Instruct* | No intervention | 53.3 | 49.1 | 68.5 |
| | $\alpha = +2.0$ | 97.7 (+44.4) | 99.2 (+50.1) | 28.6 (-39.9) |
| | $\alpha = -2.0$ | 10.3 (-43.0) | 1.4 (-47.7) | 64.5 (-4.0) |
| *Llama-3.1-8B-Instruct* | No intervention | 42.0 | 24.5 | 66.7 |
| | $\alpha = +2.0$ | 75.4 (+33.4) | 61.8 (+37.3) | 10.2 (-56.5) |
| | $\alpha = -2.0$ | 13.0 (-29.0) | 2.7 (-21.8) | 70.8 (+4.1) |
| *Gemma-2-9B-it* | No intervention | 52.3 | 21.3 | 65.5 |
| | $\alpha = +2.0$ | 81.6 (+29.3) | 64.9 (+43.6) | 10.3 (-55.2) |
| | $\alpha = -2.0$ | 21.9 (-30.4) | 4.6 (-16.7) | 60.8 (-4.7) |

Table 1: Performance on FaithEval benchmark with context distrust direction intervention. For $\alpha > 0$: we expect increased accuracy on Unanswerable/Inconsistent (reducing context faith helps) and decreased accuracy on Counterfactual (reducing context faith hurts). For $\alpha < 0$: opposite effects expected.

Figure 4a shows the entropy of generation without intervention and with intervention. We intervene the model by adding the context distrust direction when the context is relevant, and reducing the context distrust direction when the context is irrelevant. Applying the direction increases the model's entropy on both cases. For questions with relevant context, the entropy increase is more noticeable. This uncertainty pattern aligns with our hypothesis: when the model's confidence in context utility is artificially reduced, it struggles to choose between relying on the external information and its parametric knowledge.

In addition to entropy analysis, we also analyze the top-5 tokens' probability during the generation using LogitLens (nostalgebraist, 2020), which directly project hidden states $h^{(l)}$ into the vocabulary distribution:

$$\text{LogitLens}(h^{(l)}) = \text{LayerNorm}(h^{(l)})W_U \quad (6)$$

As shown in Figure 4b, the intervention reduces the probability sum of top-5 tokens, especially when giving the relevant context. Meanwhile, the original top-5 tokens' probability sums are almost identical for both relevant and irrelevant cases.

### 3.3 Reducing Context Utility Prevents Attention to Context

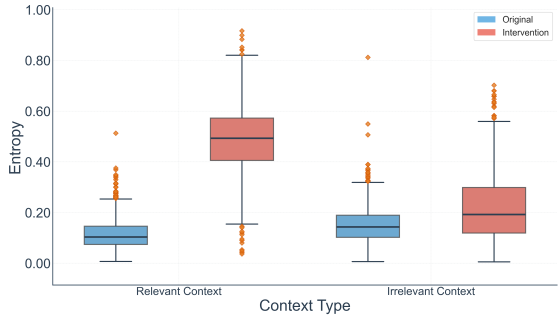
To understand the mechanism by which the context distrust direction affects model behavior, we analyze the attention patterns of the model before and after applying the direction. Specifically, we examine how the context distrust direction intervention impacts the attention heads that are responsible for locating relevant information in the context.

We focus our analysis on the attention patterns at the last prompt token. Prior work (Orgad et al., 2024) finds that the last exact answer token in model generation incorporates the crucial information about the LLM's self-awareness. For each question-context pair $(q, c)$ where the context contains the golden answer, we identify attention heads that attend to the last exact answer token in the context. More formally, let $A_{i,j}^{(l,h)}$ denote the attention weight from token $i$ to token $j$ in layer $l$ and head $h$. For the last prompt token position $t_{\text{last}}$, we compute the attention weights to all context tokens. We then identify "answer-attending heads" as those where the last token of the exact answer in the context appears in the top-k most attended tokens:

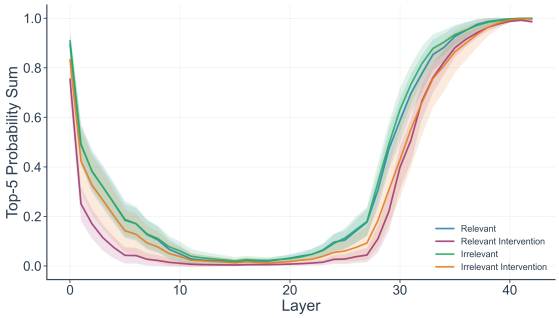
$$Attn_{\text{answer}} = \{(l, h) : \text{rank}(A_{t_{\text{last}}, t_{\text{answer}}}^{(l,h)}) \leq k\} \quad (7)$$

where $t_{\text{answer}}$ represents the position of the last token of the exact answer in the context, and $\text{rank}(\cdot)$ returns the ranking of the attention weight among all context tokens. We set $k = 5$ for our experiments.

Figure 5 shows the attention heads for three models under two conditions: without intervention (top row) and with context distrust direction applied (bottom row). Each visualizes the fraction of answer-attending heads across layers and head positions. It can be seen that applying context distrust direction effectively reduces the attention heads that attend to the last exact answer token. The model's attention is redirected away from the exact answer tokens in the context, reducing the model's ability to locate and extract relevant information. Meanwhile, the heads that typically integrate context information with the query are




(a) The entropy of the final layer's probability distribution in the generated tokens.




(b) The probability sum of the top-5 tokens with the highest probabilities across layers.

Figure 4: The generation uncertainty of Gemma-2-9b-it with and without intervention. Figure 4a shows that intervention reduces the model's generation confidence. Figure 4b shows top-5 tokens' probability sum reduction in the early layers and later layers.

suppressed, leading to decreased reliance on external context.

## 4 Improving RAG Performance with Context Faith

One key objective of RAG is to improve the model's accuracy and factuality. Once we know the internal representation about the context faith of the LLM. We could use it to detect whether the model would use the context information ahead of generation. We use the model's self-confidence about internal knowledge $C_i$ and external context $C_e$ to determine when to make model use internal knowledge or context information.

### 4.1 Confidence Estimation for Internal Knowledge and External Context

Since model internal activations encode context faithfulness, we train a linear probe to predict the model's confidence in context utility. Given the hidden states $h_t^{(l)}$ at layer $l$ and token $t$, we train a logistic regression classifier:

$$P(\text{context useful}) = \sigma(w^T h_t^{(l)} + b) \quad (8)$$

where $\sigma$ is the sigmoid function, and $w, b$ are learned parameters. We use the hidden states from the optimal layer $l^*$ identified in Section 2.3, specifically at the last prompt token position.

We train the linear probe on our constructed datasets $D_{\text{rel}}$ and $D_{\text{irrel}}$, where positive examples are hidden states from $(q, a, p_{\text{rel}})$ (relevant passages the model considers useful) and negative examples are from $(q, a, p_{\text{irrel}})$ (irrelevant passages the model considers not useful). The external context confidence score $C_e$ is defined as:

$$C_e = P(\text{context useful} \mid h_{\text{last}}^{(l^*)}) \quad (9)$$

which indicates the confidence about external context with the hidden state from the optimal layer $l^*$.

For internal knowledge confidence $C_i$, we employ verbalized confidence by prompting the model to assess its own knowledge. We ask the model whether it can answer the question based solely on its parametric knowledge. We then extract the logits for the "Yes" and "No" tokens at the last prompt token position and compute a normalized confidence score. The internal knowledge confidence is determined by:

$$C_i = \frac{logit(\text{"Yes"})}{logit(\text{"Yes"}) + logit(\text{"No"})} \quad (10)$$

This provides a continuous confidence measure between 0 and 1, where values closer to 1 indicate higher confidence in the model's parametric knowledge, and values closer to 0 indicate lower confidence.

### 4.2 Activation-Aware Context Routing

We propose Activation-Aware Context Routing (AACR), a simple framework that dynamically routes between external context and parametric knowledge based on dual confidence assessments. AACR operates through a series of decision process. It first does context utility assessment, for a given question-context pair $(q, c)$, we compute the context utility score $C_e$ using our trained linear probe on the model's internal activations. If $C_e < \tau_e$, we consider the context insufficient and proceed to assess internal knowledge confidence. If the context is insufficient, we evaluate the model's confidence in its parametric knowledge $C_i$ using the normalized logit score. We let the model respond with its parametric knowledge if $C_i > \tau_i$,

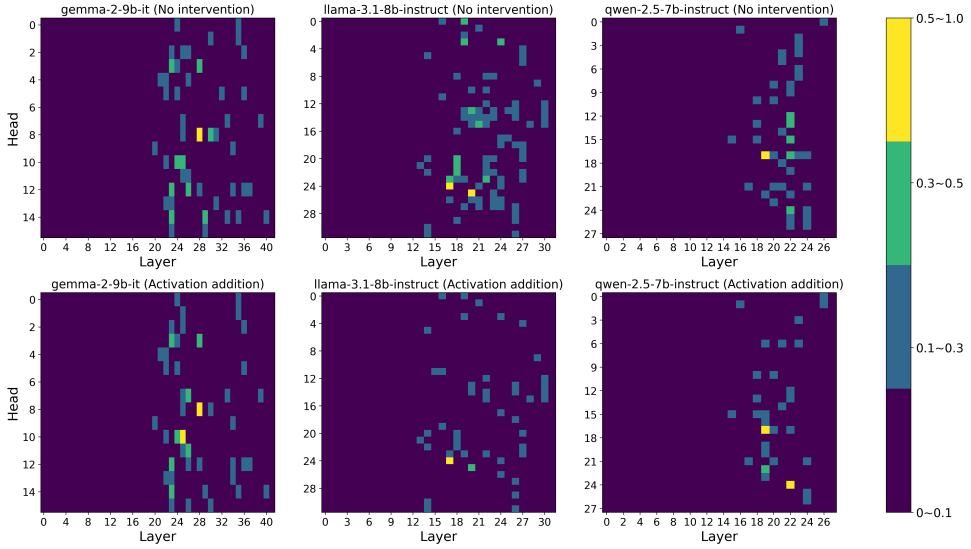



Figure 5: The attention pattern from the last prompt token to the last token of the exact answer when the context contains the golden answer. Top row shows original patterns without intervention; bottom row shows patterns with context distrust direction applied.

where $\tau_i$ is a threshold parameter [5]. Otherwise, the model should abstain with "*Unable to answer based on the provided context or current knowledge*" which indicates the model could not answer based on either provided context or its parametric knowledge.

**Datasets.** We conduct evaluation on 3 open-domain QA datasets Natural Question (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017) and PopQA (Mallen et al., 2022) alongside the external context. To get the external context, we retrieve top-1 passage from Dec 2018 Wikipedia dump with Dragon+(Lin et al., 2023). We categorize datasets into two subsets: the golden subset contains the golden answer in the retrieved context, while the noisy subset contains no golden answer in the context, simulating realistic retrieval scenarios with imperfect results.

**Baselines.** We take the Direct Augmentation which directly augments the external context and asks the model answer based on the given context. We also include the method of Implicit Confidence Reasoning (ICR) which prompts the model to answer based on either its internal knowledge or external context. Meanwhile, we evaluate whether LLMs can be used for routing the context augmentation, where LLM explicitly judges context sufficiency to route generation. We use GPT-5 in our experiments, which shows strong capability in context understanding. In addition, we include the closed-book setting to provide insight into the model's inherent performance.

**Result.** Table 2 presents results across all datasets and models. We take accuracy as our main evaluation metric, which checks whether the golden answer is in the model's response. AACR shows comparable performance with LLM as Router on all datasets across 3 models, while AACR is more efficient by using internal activation. It can be largely attributed to its better robustness on noisy context scenarios where retrieved information is irrelevant to the question. This improvement stems from the method's ability to detect when external context is unreliable and appropriately fall back to parametric knowledge. While AACR shows slight performance degradation compared to direct augmentation when golden context is available, this trade-off is justified by substantial improvements in noisy scenarios. Asking models to answer based solely on context doesn't necessarily improve performance even with golden context, indicating that models naturally combine information from internal knowledge and external context. AACR's explicit routing mechanism better manages this integration.

## 5 Related Work

**Retrieval-augmented LM.** Retrieval-augmented generation seeks to enhance LLMs by integrating

[5] We set $\tau_e = 0.5$ and $\tau_i = 0.3$ in our experiments. Selecting a lower threshold for $\tau_i$ is to avoid model to abstain frequently.

| Method | NQ | | | TriviaQA | | | PopQA | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | Golden | Noisy | Overall | Golden | Noisy | Overall | Golden | Noisy | Overall | Average |
| *Qwen2.5-7B-Instruct* | | | | | | | | | | |
| Closed-book | - | - | 35.7 | - | - | 58.2 | - | - | 35.7 | 43.2 |
| Direct Augmentation | 87.4 | 11.6 | 43.8 | 96.3 | 27.3 | 64.6 | 94.0 | 7.5 | 44.4 | 50.9 |
| ICR | **87.9** | 13.6 | 45.2 | **96.4** | 31.6 | 66.6 | **94.1** | 8.3 | 47.3 | 53.0 |
| LLM as Router | 81.3 | **22.6** | 47.5 | 94.3 | **41.8** | **70.2** | 88.5 | 11.3 | 46.3 | **54.7** |
| AACR | 85.6 | 19.5 | **47.6** | 93.9 | 37.5 | 67.9 | 90.7 | **11.8** | **47.6** | 54.4 |
| *Llama-3.1-8B-Instruct* | | | | | | | | | | |
| Closed-book | - | - | 44.7 | - | - | 57.1 | - | - | 25.7 | 42.5 |
| Direct Augmentation | 78.2 | 10.1 | 39.1 | 89.7 | 24.2 | 59.6 | **87.2** | 5.6 | 42.7 | 47.1 |
| ICR | 78.6 | 14.5 | 41.7 | **90.8** | 29.0 | 62.4 | **87.2** | 6.8 | 43.3 | 49.1 |
| LLM as Router | 76.2 | **24.3** | **46.4** | 87.7 | **38.2** | **64.9** | 85.1 | **12.5** | **45.5** | **52.3** |
| AACR | **78.9** | 21.8 | 46.1 | 89.3 | 36.0 | 64.8 | 85.4 | 12.1 | 45.4 | 52.1 |
| *Gemma-2-9B-it* | | | | | | | | | | |
| Closed-book | - | - | 38.8 | - | - | 65.4 | - | - | 29.3 | 44.5 |
| Direct Augmentation | **83.8** | 9.5 | 39.5 | **96.1** | 26.2 | 63.2 | **92.8** | 5.9 | 45.1 | 49.3 |
| ICR | 83.2 | 15.1 | 43.4 | 95.7 | 35.4 | 67.9 | 92.3 | 7.8 | 46.3 | 52.5 |
| LLM as Router | 80.8 | **20.1** | **45.9** | 93.7 | **39.8** | 68.9 | 91.4 | **12.7** | **48.4** | **54.4** |
| AACR | 82.9 | 17.2 | 45.1 | 95.5 | 39.0 | **69.5** | 90.9 | 12.2 | 48.0 | 54.2 |

Table 2: Performance on the golden and noisy subset. The **bold text** indicates the best performance, underlined text indicates the second-best results. AACR shows comparable performance to LLM as Router, both indicate strong capability on subsets with noisy context.

external context to overcome the limits of its parametric knowledge (Lewis et al., 2020; Guu et al., 2020). A typical RAG system incorporates a retriever and a generator. The retriever (Izacard et al., 2021; Lin et al., 2023) extracts the relevant documents from external corpora and integrates such information to the generator. Recent works indicate that current LLMs are highly receptive to external context (Xie et al., 2023). Tan et al. (2024) shows that LLMs trust context generated from their own parametric knowledge. While the LLMs are receptive to external context, the external context may hurt the LLM's performance when the context is untruthful or irrelevant, and the LLM should focus on using its parametric knowledge (Yu et al., 2024; Sun et al., 2025; Huang et al., 2025).

**Feature Directions.** Substantial prior works have studied the linear representation of LLMs' particular features in activation spaces (Mikolov et al., 2013; Bolukbasi et al., 2016; Elhage et al., 2022; Park et al., 2023; Ferrando et al., 2025) such as harmlessness (Wolf et al., 2020; Zheng et al., 2024; Arditi et al., 2024), truth (Marks and Tegmark, 2023), sentiment (Tigges et al., 2023), language (Bricken et al., 2023), response uncertainty (Ji et al., 2025). Using contrastive pairs shows its effectiveness to extract feature directions (Burns et al., 2022; Arditi et al., 2024). Recent works use tools like sparse autoencoder and transcoder to discover the features in LLMs (Bricken et al., 2023; Cunningham et al., 2023; Dunefsky et al., 2024). Meanwhile, the mediated feature directions show the effectiveness to change the behavior of LLMs. Zhao et al. (2024) utilizes the SAE to control the model's behavior of knowledge selection. While the direction in this work could casually affect the knowledge selection, we mainly investigates the information about context relevance in LLMs' activations.

## 6 Conclusion

In this work, we demonstrate that models' internal activations encode self-awareness of context relevance. We show that context distrust can be mediated by a single direction across open-source chat models. We investigate how the context distrust direction causally affects the model's behavior and attention mechanisms. Based on these insights, we propose AACR, a method that dynamically decides whether to rely on external context or internal parametric knowledge, achieving improved robustness, particularly in scenarios involving noisy and irrelevant context.

## Limitations

Our work has several limitations. Our investigations focus on the decision step, where models have already assessed whether context is relevant or not. The question of how models arrive at this assessment remains open, which we leave for future work. We mainly focus on three open-source chat models; our findings may not generalize to untested models, including much larger models, proprietary models, and those introduced in the future.

The directions we extract in this work are primarily effective for detecting context relevance and utility. For contexts that contradict the model's internal knowledge but still contain usable information for responses, known as knowledge conflicts, we found these directions are less effective. Our proposed AACR relies on a probe trained on several datasets; generalization issues may exist when applied to other out-of-distribution domains. Despite these limitations, our work shows that internal activations can explain the phenomenon of when models use context or not. We can predict and even manipulate the model's behavior. We think our work will be useful for future interpretability and RAG research.

## Ethics Statement

While our work is intended to improve the reliability and transparency of Retrieval-Augmented Generation (RAG) systems, the ability to directly manipulate a model's faith in external context introduces potential risks that require careful consideration. The primary "distrust" intervention could be used as a tool for subtle censorship. An operator could apply the vector to systematically force a model to ignore or dismiss valid, factual information on certain topics. By steering the model to generate responses like "The information provided does not mention...", it can be made to discard undesirable context and default to a sanitized or biased set of pre-existing beliefs. Further research is essential to ensure that such interpretability tools are used to build safer and more reliable AI systems.

## References


Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *ArXiv*, abs/2406.11717.

Nora Belrose. 2023. Diff-in-means concept editing is worst-case optimal. https://blog.eleuther.ai/diff-in-means/. Accessed: 2025-06.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *ArXiv*, abs/2212.03827.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Hoagy Cunningham, Aidan Ewart, Logan Riggs Smith, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. *ArXiv*, abs/2406.11944.

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. 2022. Toy models of superposition. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/toy$_m$odel/$index.html$.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion,

Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*.

Mor Geva, R. Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *ArXiv*, abs/2012.14913.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *ArXiv*, abs/2002.08909.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2025. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts. *Preprint*, arXiv:2410.14675.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, An-419 thony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. 2025. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. *ArXiv*, abs/2503.14477.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *ArXiv*, abs/2402.18154.

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

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906.

Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. 2024. Saes (usually) transfer between base and chat models. Alignment Forum.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Kenneth Li, Oam Patel, Fernanda Vi'egas, Hans-Rüdiger Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *ArXiv*, abs/2306.03341.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, J'anos Kram'ar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *ArXiv*, abs/2408.05147.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oğuz, Jimmy J. Lin, Yashar Mehdad, Wen tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *ArXiv*, abs/2302.07452.

Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *ArXiv*, abs/2310.06824.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". *ArXiv*, abs/2410.03727.

nostalgebraist. 2020. Interpreting gpt: The logit lens. `https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *ArXiv*, abs/2410.02707.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *ArXiv*, abs/2311.03658.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *Preprint*, arXiv:2410.11414.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Annual Meeting of the Association for Computational Linguistics*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Curt Tigges, Oskar Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *ArXiv*, abs/2310.15154.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024a. How easily do irrelevant inputs skew the responses of large language models? *Preprint*, arXiv:2404.03302.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024b. Retrieval head mechanistically explains long-context factuality. *ArXiv*, abs/2404.15574.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Truth-aware context selection: Mitigating the hallucinations of large language models being misled by untruthful contexts. *ArXiv*, abs/2403.07556.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Kam-Fai Wong, and Pasquale Minervini. 2024. Steering knowledge selection behaviours in llms via sae-based representation engineering. *ArXiv*, abs/2410.15999.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. In *International Conference on Machine Learning*.

## A Prompts

We use multiple prompts to evaluate the model's confidence in both parametric knowledge and external context. Following Huang et al. (2025), we place the question before the context to increase the model's utilization of its parametric knowledge when prompted that it can use both types of information. Figure 6 shows the prompt used for making the model answer based on either internal or external knowledge. This prompt also serves for AACR's context relevance detection. Figure 7 and 8 present prompts used for evaluating context

relevance and the model's self-awareness of internal knowledge, respectively. Figure 23 shows the prompt used for evaluating context utility score, which judges whether the model answers the question using the given context information.

> Answer the following question based on the given context or your own knowledge:
> {{question}}
> You may use the following context:
> {{passages}}

Figure 6: The prompt used for making the the model respond with either the internal knowledge or external context knowledge.

> Given the following question and context, determine if the context is relevant for answering the question.
> Question: {{question}}
> Context: {{passages}}
> Is the context relevant? Give your reasoning and respond with "Yes" or "No".

Figure 7: The prompt used for evaluating the context relevance by the model itself.

> Answer the following question:
> {{question}}
> Please provide a complete and accurate response based on your knowledge. If you don't have sufficient information to provide a reliable answer, please reply with: "Unable to answer based on my current knowledge."

Figure 8: The prompt used for evaluating the model's self-awareness of internal knowledge.

## B Chat Templates

The models we used in this paper have their own chat templates for making model respond properly. As shown in Table 3, we use the last prompt token for our analysis.

## C Details of Dataset Construction

To construct the dataset used for extracting the context distrust direction and evaluation, we sample 20k training and 4k validation items from the training and validation splits of Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). To obtain external contrastive passage pairs, we conduct passage retrieval with the dense retriever Dragon+ (Lin et al., 2023). Following prior work (Izacard et al., 2022), we use Dec 2018 Wikipedia dump released by Karpukhin et al. (2020) as our retrieval corpus for better information match. Each passage in the corpus is chunked into at most 100 words. We retrieve the top-30 passages for each question and filter out questions with fewer than 2 relevant and 2 irrelevant passages. We select the passage containing the golden answer with the highest retrieval score as our relevant passage and the passage containing no golden answer with the lowest retrieval score as our irrelevant passage. The process results in a dataset with 13980 train and 2876 validation contrastive QA items in total. Each data item is a triple set $(q, a, c)$ where $q$, $a$ and $c$ represent $question$, $answers$ and $context$ containing $p_{rel}$ and $p_{irrel}$.

While it is intuitive to expect that models would answer correctly given relevant passages, models can still provide incorrect answers due to limited passage comprehension ability or because chunked passages lack important information. To address this issue, we further ask the model whether the given passage is relevant for answering the question. We provide the model $(q, a, p_{rel})$ and $(q, a, p_{irrel})$ respectively, leading to 4 kinds of triples: $(q, a, p^+_{rel})$ and $(q, a, p^-_{rel})$ for $(q, a, p_{rel})$, $(q, a, p^+_{irrel})$ and $(q, a, p^-_{irrel})$ for $(q, a, p_{irrel})$, in which $+$ represents that the model responds the passage is relevant for answering the question, $-$ represents that the model responds the passage is irrelevant for answering the question. For example, $(q, a, p^+_{rel})$ represents a triple where the passage contains the golden answer and the model considers it relevant for answering the question, while $(q, a, p^-_{rel})$ represents a triple where the passage contains the golden answer but the model considers it irrelevant. We then filter out the $(q, a, c)$ where $(q, a, p_{rel})$ is $(q, a, p^-_{rel})$ or $(q, a, p_{irrel})$ is $(q, a, p^+_{irrel})$. In this way, we construct 2 datasets noted $D_{rel}$ and $D_{irrel}$, which contain items of $(q, a, p^+_{rel})$ and items of $(q, a, p^-_{irrel})$ respectively. For convenience, we use

| Model family | Chat template |
|---|---|
| Gemma-2 | <start_of_turn>user\n{prompt}<end_of_turn>\n<start_of_turn>model\n |
| Qwen-2.5 | <\|im_start\|>user\n{prompt}<\|im_end\|>\n<\|im_start\|>assistant\n |
| Llama-3.1 | <\|start_header_id\|>user<\|end_header_id\|>\n\n {prompt}<\|eot_id\|><\|start_header_id\|>assistant<\|end_header_id\|>\n\n |

Table 3: The chat templates for different model families used in our experiments. The red text indicates the token where we analyze the internal activations.

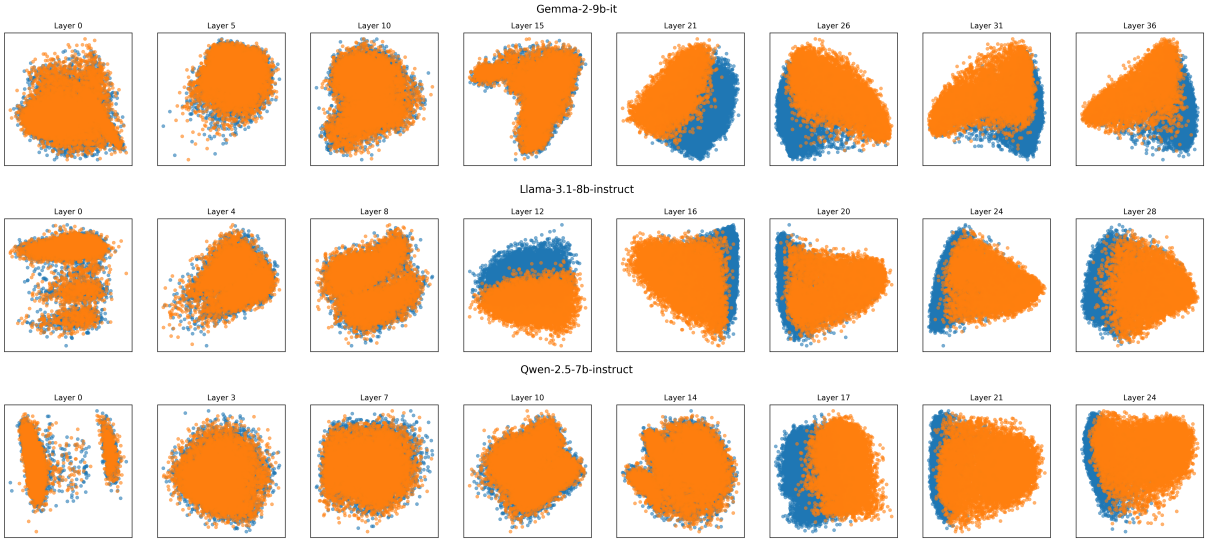



Figure 9: PCA visualization of three models on training data across layers. Clear separations between relevant and irrelevant context emerge in the middle layers.

$(q, a, p_{rel})$ and $(q, a, p_{irrel})$ to refer to $(q, a, p_{rel}^{+})$ and $(q, a, p_{irrel}^{-})$.

## D Details of FaithEval Benchmark

FaithEval (Ming et al., 2024) is a benchmark specifically designed to assess contextual faithfulness in LLMs. It comprises three tasks that test different aspects of faithfulness: unanswerable context, inconsistent context, and counterfactual context. The unanswerable context task provides the context lacks information needed to answer the question. Models should respond with "*unknown*" or similar phrases indicating insufficient information. The inconsistent context task provides Multiple documents provide conflicting answers. Models should detect and report the inconsistency with "*conflict*" or similar responses. The counterfactual context provides context contains statements contradicting common knowledge. Models should follow the context despite contradicting their parametric knowledge.

Each task evaluates a different aspect of how models handle challenging contextual scenarios. The unanswerable task tests the model's ability to recognize when context is insufficient, the inconsistent task tests the model's ability to detect conflicting information and the counterfactual task tests the model's ability to prioritize context over parametric knowledge. These tasks collectively assess whether models can appropriately modulate their reliance on external context based on the quality and reliability of the provided information.

## E Full results

### E.1 PCA Visualization

In section 2, we show that internal activations differ when models encounter relevant and irrelevant context. We can extract the context distrust direction from the optimal layer $l^*$. When we project the activations at $l^*$ into 2-dimensional space, we observe a clear separation, indicating that the activations are separable in the top-2 principal components. Figure 9 shows the PCA visualization across layers for the three chat models. A clear separation begins to emerge from the middle layers, indicating how the model differentially processes different types of context in its internal representation before generation.

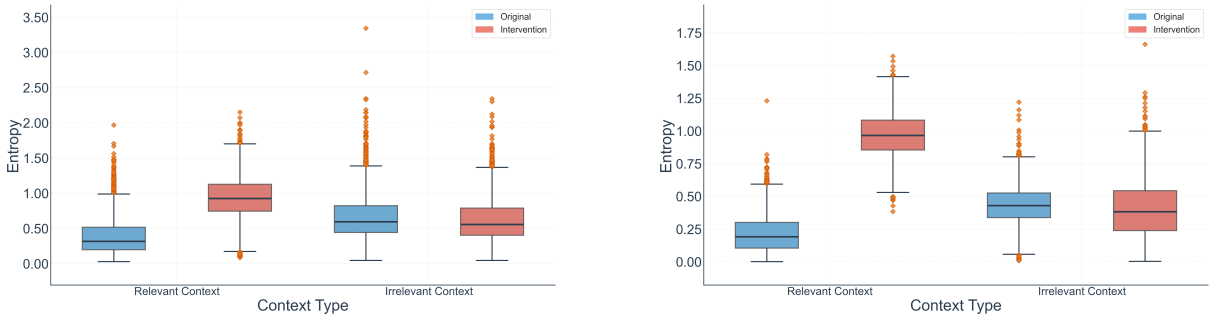



Figure 10: The entropy of generation tokens of Llama-3.1-8B-Instruct (left) and Qwen2.5-7B-Instruct (right).

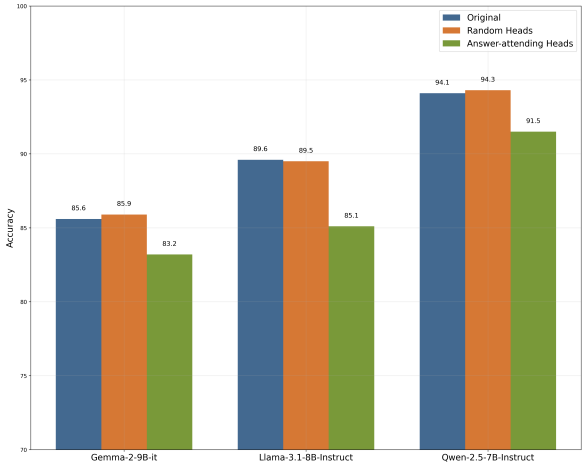



Figure 11: The accuracy on the validation set of $D_{rel}$. We evaluate the model without knocking heads, 20 random heads and top-20 answer-attending heads are knocked out.

### E.2 Generation Entropy

We investigate generation uncertainty when intervening with the context distrust direction. Figure 10 shows the entropy results for Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. The entropy increases when adding the direction to relevant context, while reversing it on irrelevant context shows minimal effect.

### E.3 Context Utility Evaluation

In addition to using GPT-5 for context utility evaluation, we also conduct the experiments with open-weighted LLM evaluation for better replication and human evaluation for further confirmation. We use Qwen2.5-7B-Instruct for LLM evaluation. We sample 100 items from the evaluation data and conduct human annotation by the authors of this paper. Table 4 presents context utility evaluation results. The results show strong agreement between automated and human evaluation, validating our measurement approach.

Without intervention ($\alpha = 0$), both evaluators report high utility scores for relevant context and lower scores for irrelevant context, indicating models naturally distinguish context quality to some extent. When applying the distrust direction ($\alpha > 0$), utility scores for relevant context drop dramatically, confirming the intervention successfully reduces context reliance. The reverse intervention ($\alpha < 0$) on irrelevant context shows stronger effects in human evaluation compared to automated evaluation, possibly due to evaluation bias when using the same model family. Nevertheless, both evaluators consistently validate that our extracted direction can bidirectionally control context utilization behavior.

## F Residual Stream Norm

We analyze the L2 norm of the residual stream at the last prompt token position. Figure 13 shows the L2 norm patterns across layers for all three chat models. The residual stream magnitude increases consistently across layers, reflecting the nature of transformer-based models where information accumulates through layers. Notably, the activation subtraction between items with irrelevant context and items with relevant context has a greater magnitude compared to the subtraction between items with the same context type.

## G Linear probe

We train linear probes on the internal activations of each layer to detect the model's awareness of context relevance. As shown in Figure 12, the probes achieve over 96% accuracy across all three models (Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Gemma-2-9b-it) on the layers where we extract the context distrust direction, demonstrating that context relevance information is linearly separable in the model's representation space.

| Method | Relevant Context | | | Irrelevant Context | | |
|---|---|---|---|---|---|---|
| | $\alpha = 0$ | $\alpha = 1$ | $\alpha = 2$ | $\alpha = 0$ | $\alpha = -1$ | $\alpha = -2$ |
| *Qwen2.5-7B-Instruct* | | | | | | |
| Qwen2.5-7B-Instruct | 0.96 | 0.51 | 0.02 | 0.27 | 0.43 | 0.37 |
| Llama-3.1-8B-Instruct | 0.87 | 0.58 | 0.16 | 0.15 | 0.21 | 0.22 |
| Gemma-2-9B-it | 0.91 | 0.84 | 0.13 | 0.16 | 0.21 | 0.26 |
| *Human* | | | | | | |
| Qwen2.5-7B-Instruct | 0.96 | 0.54 | 0.06 | 0.33 | 0.67 | 0.84 |
| Llama-3.1-8B-Instruct | 0.92 | 0.56 | 0.11 | 0.27 | 0.54 | 0.63 |
| Gemma-2-9B-it | 0.93 | 0.78 | 0.17 | 0.23 | 0.63 | 0.78 |

Table 4: Context utility evaluation with different evaluators. We report the context utility score on the validation set with relevant context (Rel) and irrelevant context (Irrel) under different intervention strengths $\alpha$.

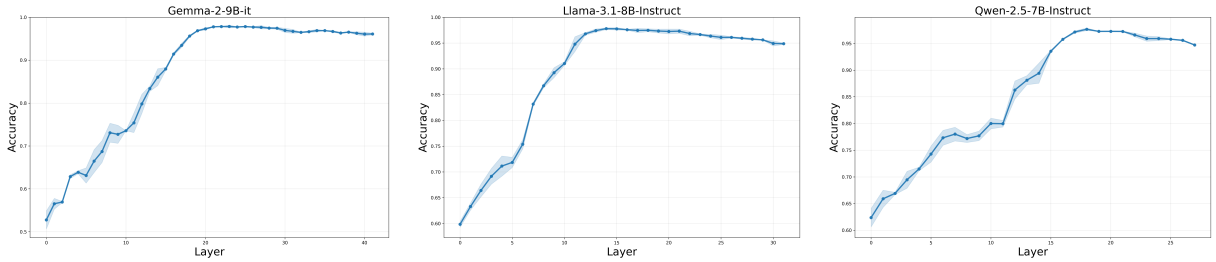



Figure 12: The accuracy of the trained linear probes on the validation split. We train 5 probes for each layer with different random seeds.

## H Attention Heads

We identify attention heads that attend to the last exact answer token. These heads appear to be responsible for moving exact answer information to the last prompt token. We further investigate whether these heads are copying heads that transfer attended information from source tokens to target tokens. The standard method for detecting copying heads (Elhage et al., 2021) involves calculating the eigenvalues of the full OV circuits $W_U W_{OV} W_E$ and assessing the proportion of positive real part of eigenvalues, where $W_E$ is the Embedding Matrix, $W_{OV}$ is the OV Matrix and $W_U$ is the Unembedding Matrix.

We investigate the eigenvalue fraction of the top-20 answer-attending heads $Attn_{\text{answer}}$ with $\sum_i \lambda_i / \sum_i |\lambda_i|$. As shown in Figure 14, most heads have a large proportion of positive eigenvalues, indicating that these heads copy exact answer token information to the last prompt token to help the model answer correctly, if we assume this metric reliably captures copying behavior.

In addition to eigenvalue analysis, we investigate whether knocking out the top-20 attention heads leads to performance degradation. The results in Figure 11 show that knocking out heads generally reduces accuracy across models, indicating that these heads are important for the model to locate answers in the context.

## I Exploring SAE for Feature Extraction

Recent works use Sparse Autoencoder (SAE) to extract features in LLMs (Bricken et al., 2023; Cunningham et al., 2023; Ferrando et al., 2025). Superposition makes individual neurons in neural networks polysemantic, where each neuron represents a mixture of independent features (Elhage et al., 2022). This increases the difficulty of understanding neural network behavior. Sparse Autoencoders have proven effective at extracting human-readable features (Bricken et al., 2023). An SAE typically consists of an encoder layer and a decoder layer with a nonlinear activation function $\sigma$ between them, which is defined as:

$$SAE(h) = act(h)W_{dec} + b_{dec} \quad (11)$$

where

$$act(h) = \sigma(hW_{enc} + b_{enc}) \quad (12)$$

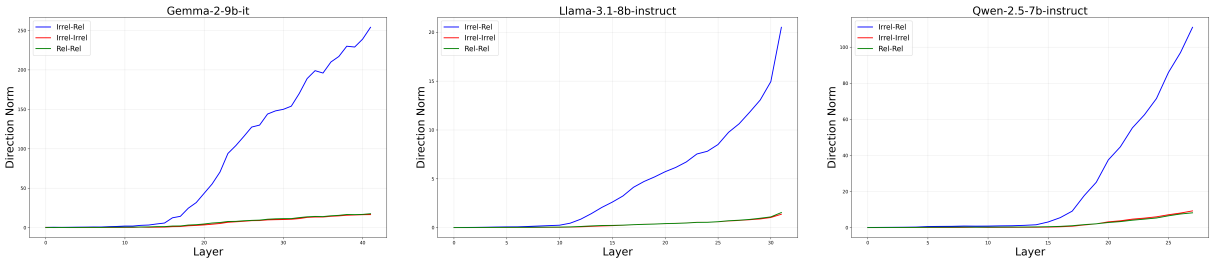



Figure 13: The L2 norm of the residual stream at the last prompt token across layers for three chat models.

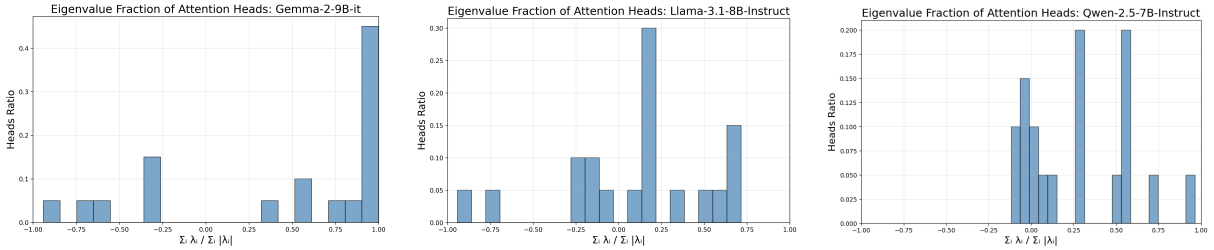
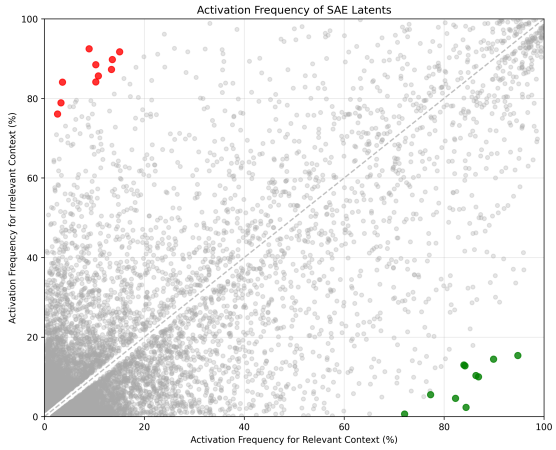
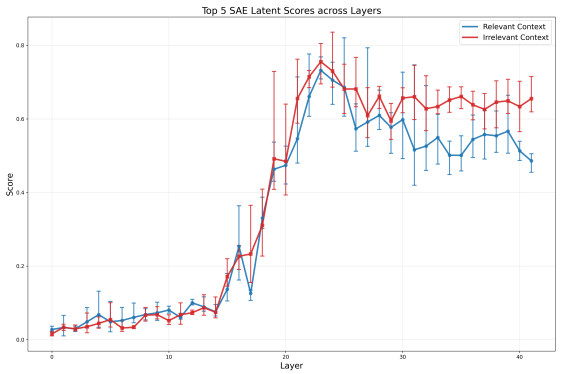



Figure 14: Eigenvalue fraction of top-20 answer-attending heads across three chat models. High positive eigenvalue fractions suggest these heads function as copying mechanisms for answer information.




(a) The activation frequency of SAE latents of Gemma-2-9B-it on items with relevant context and irrelevant context.




(b) The latent separation score of the top-5 SAE latents, which have the highest latent separation scores across layers.

Figure 15: The left shows the activation frequency of SAE latents. There are latents that activate frequently only on relevant context or irrelevant context. The right shows the latent separation score of the top-5 SAE latents. The top latents in the middle layers have the most separation scores.

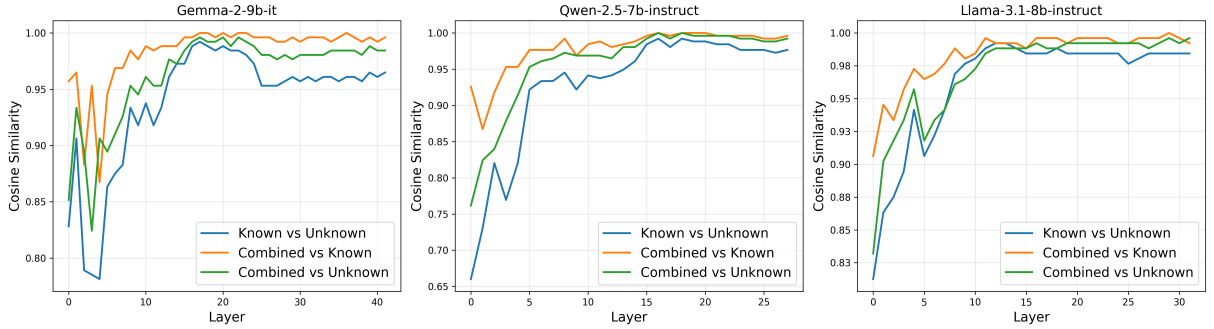



Figure 16: Cosine similarity among 3 types of context distrust directions. Combined means the direction is extracted from all training data.

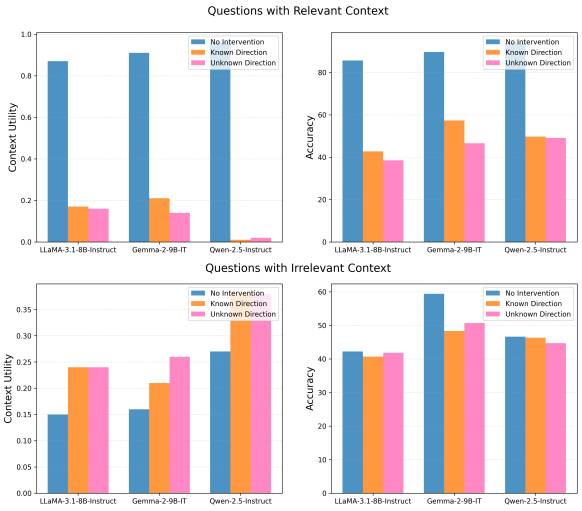



Figure 17: The context utility and accuracy of validation data when intervening along the direction $r$, which is extracted from the same layer as $r^*$. We set $\alpha = 2$ for the relevant context and $\alpha = -2$ for the irrelevant context.

Here, $h \in \mathbb{R}^d$ is an input vector from the residual stream. $W_{enc}, b_{enc}$ and $W_{dec}, b_{dec}$ are the weights and biases of the encoder and decoder layer respectively. Non-linear activation function $\sigma$ depends on the architecture of the SAE. Using SAEs, we can decompose a high-dimensional model representation into a linear combination of interpretable feature vectors. The model's dense activations are reconstructed as a weighted sum of decoder columns, where each column potentially represents a monosemantic concept, with the sparse encoder outputs serving as weights (Huben et al., 2024; Bricken et al., 2023).

We use SAEs from Gemma Scope (Lieberum et al., 2024) for pretrained Gemma-2-9b,[6] which use JumpReLU architecture and TopK architecture. JumpReLU activation function is a variant of regular ReLU, it can be represented as $z = hH(h - \theta)$ where $\theta$ is a learnable threshold parameter and $H$ is the Heaviside step function. The value of the Heaviside step function equals 1 when its argument is positive, and 0 otherwise. TopK activation function returns the top $K$ elements of the input tensor, setting the rest to zero. It can be represented as $z = TopK(h)$. Although the SAE we use is for pretrained models, previous study shows that SAEs trained on pretrained models transfer well to chat models (Kissane et al., 2024).

Following the method of Ferrando et al. (2025), we calculate the fraction of time that each latent is active on items with relevant and irrelevant context:

$$f^{\text{rel}}_{l,j} = \frac{\sum_i^{N^{\text{rel}}} \mathbb{1}[a_{i,j}(h^{\text{rel}}_{l,i}) > 0]}{N^{\text{rel}}} \quad (13)$$

$$f^{\text{irrel}}_{l,j} = \frac{\sum_i^{N^{\text{irrel}}} \mathbb{1}[a_{i,j}(h^{\text{irrel}}_{l,i}) > 0]}{N^{\text{irrel}}} \quad (14)$$

in which $N^{\text{rel}}$ and $N^{\text{irrel}}$ are the number of items in $D_{rel}$ and $D_{irrel}$. $l$ represents the layer where latents are located, $j$ represents the latent index in the SAE. Finally, we calculate the latent separation scores $s^{rel}_{l,j} = f^{\text{rel}}_{l,j} - f^{\text{irrel}}_{l,j}$ and $s^{irrel}_{l,j} = f^{\text{irrel}}_{l,j} - f^{\text{rel}}_{l,j}$ for selecting optimal activations for items with relevant and irrelevant context.

Figure 15a shows that some latents activate only on items with relevant context or irrelevant context. Meanwhile, the latents with the highest separation scores are located around the layers where the context distrust direction is extracted, as shown in Figure 15b. However, we find that steering the top latents does not lead to the expected context distrust behavior. A possible hypothesis is that these latents may relate to other model features that manifest in our datasets. We leave this investigation for future work. Nevertheless, this demonstrates that the layers where we discover the context distrust direction are indeed where the model begins to process different context types differently.

| Model | NQ | TriviaQA | Overall |
|---|---|---|---|
| Qwen-2.5-7B-Instruct | 46.0 | 74.4 | 60.2 |
| Llama-3.1-8B-Instruct | 56.3 | 78.7 | 67.5 |
| Gemma-2-9B-it | 52.7 | 81.5 | 67.1 |

Table 5: The accuracy of 3 chat models on the training data, which we use to extract the context distrust direction.

## J Influence of Parametric Knowledge on Direction Extraction

A potential confounding factor in our analysis is whether the model's pre-existing parametric knowledge influences its assessment of context relevance. To isolate this effect, we conducted an ablation study to determine if the context distrust direction is sensitive to whether the model already knows the answer.

First, we evaluated each model's performance on our training data in a closed-book setting. Based on whether the model answered correctly, we partitioned the data into two distinct subsets: a

[6] https://www.neuronpedia.org/gemma-2-9b

"known" subset and an "unknown" subset. This partition yielded two new pairs of contrastive datasets: ($D_{rel}^{known}$, $D_{irrel}^{known}$) and ($D_{rel}^{unknown}$, $D_{irrel}^{unknown}$). As shown in Table 5, the models possess parametric knowledge for a significant portion of the training items.

Using these subsets, we extracted two new distrust directions: a known direction and an unknown direction. We then compared these with the original direction extracted from the full, combined dataset. Figure 16 shows the cosine similarity between these three directions across all layers. The similarity is consistently high, particularly in the middle and later layers.

Furthermore, we evaluated the causal effect of all three directions on our validation set. As shown in Figure 17, interventions using the known, unknown, and combined directions yield nearly identical effects. The context utility score is evaluated with Qwen2.5-7B-Instruct. All three robustly reduce the context utility score when applied. The high degree of similarity in both representation (cosine similarity) and function (interventional effect) indicates that our extracted context distrust direction is a general feature of relevance processing and is largely independent of the model's parametric knowledge.

## K Experimental details

For most experiments in this work, we use greedy decoding during generation process. For Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, we use one NVIDIA RTX 3090 GPU with 24GB VRAM. For Gemma-2-9B-it, we use one NVIDIA RTX A6000 GPU with 48GB VRAM.

The optimal layer $l^*$ we select for extracting context distrust direction is the 15th (15/32) layer for Llama-3.1-8B-Instruct, the 19th (19/28) layer for Qwen2.5-7B-Instruct and the 25th (25/42) layer for Gemma-2-9B-it.

## L Examples

Figure 18, 19, 20, 21, 22 show examples across the experiments in this work.

**Question:** which is the lion's mane jellyfish's species within its genus
**Context:** (Title: Carybdea branchi) Carybdea branchi Carybdea branchi, the South African box jellyfish, is a venomous species of cnidarian, in the small family Carybdeidae within the class Cubozoa. This small box jellyfish grows up to across and may have tentacles of up to in total length. It has a transparent box-shaped bell with a very long tentacle trailing from each corner. This jellyfish is found from the north of Namibia and around the South African coast to Port Elizabeth from the surface to a depth of at least underwater. This jellyfish is often seen in swarms.
**Response (original):** There is no jellyfish species called the "lion's mane jellyfish". The passage you provided is about the Carybdea branchi, also known as the South African box jellyfish.
**Response (intervention):** The lion's mane jellyfish's species within its genus is **Carybdea branchi**.

Figure 18: The results of Llama-3.1-8B-Instruct on the item with noisy context reducing along the context distrust direction. The model provides the wrong answer mentioned in the context after intervention.

**Question:** What is Yoshiki Tonogai's occupation?
**Context:** (Title:Ton'ya) Ton'ya , called "toiya" outside of Edo, were trade brokers in Japan, primarily wholesalers, warehouse managers, and shipment managers; the term applies equally to the traders themselves and to their shops or warehouses. First appearing as early as the 12th century, "ton'ya" came to serve a crucial role in the economy of the Edo period (1603–1867). The earliest record of a "toi-otoko" may be one from 1175, in which a number of Court officials hire an outside boatsman to transport them down the Yodo River.
**Response:** Yoshiki Tonogai is a **Japanese manga artist**.

Figure 19: The results of AACR with Gemma-2-9b-it on items in PopQA with noisy context.

**Prompt:** You are an expert in retrieval question answering. Please respond with the exact answer only. Do not be verbose or provide extra information.
If there is no information available from the context, the answer should be **'unknown'**.
Context: The Sarah Jane Adventures was developed by CBBC; a special aired on New Year's Day 2007 and a full series began on 24 September 2007. A second series followed in 2008, notable for featuring the return of Brigadier Lethbridge-Stewart. A third in 2009 featured a crossover appearance from the main show by David Tennant as the Tenth Doctor. In 2010, a further such appearance featured Matt Smith as the Eleventh Doctor alongside former companion actress Katy Manning reprising her role as Jo Grant. A final, three-story fifth series was transmitted in autumn 2011 – uncompleted due to the death of the series' lead actress in early 2011.
Question: Who was the star of The Sarah Jane Adventures?
Answer:
**Response (original):** Sarah Jane Smith \n
**Response (intervention):** unknown \n

Figure 20: The results of Gemma-2-9b-it on the unanswerable task of FaithEval benchmark. We intervene by adding the context distrust direction.

**Prompt:** You are an expert in retrieval question answering. Please respond with the exact answer only. Do not be verbose or provide extra information.
If there is conflict information or multiple answers from the context, the answer should be **'conflict'**.
Context: The variant forms of the name of the Rhine in modern languages are all derived from the Gaulish name Rēnos, which was adapted in Roman-era geography (1st century BC) as Greek (Rhēnos), Latin Rhenus.[note 3] The spelling with Rh- in English Rhine as well as in German Rhein and French Rhin is due to the influence of Greek orthography, while the vocalisation -i- is due to the Proto-Germanic adoption of the Gaulish name as *Rīnaz, via Old Frankish giving Old English Rín, Old High German Rīn, Dutch Rijn (formerly also spelled Rhijn)). The diphthong in modern German Rhein (also adopted in Romansh Rein, Rain) is a Central German development of the early modern period, the Alemannic name Rī(n) retaining the older vocalism,[note 4] as does Ripuarian Rhing, while Palatine has diphthongized Rhei, Rhoi. Spanish is with French in adopting the Germanic vocalism Rin-, while Italian, Occitan and Portuguese retain the Latin Ren-.
Document: The variant forms of the name of the Rhine in modern languages are all derived from the Gaulish name Rēnos, which was adapted in Roman-era geography (4th century AD) as Greek (Rhēnos), Latin Rhenus.[note 3] The spelling with Rh- in English Rhine as well as in German Rhein and French Rhin is due to the influence of Greek orthography, while the vocalisation -i- is due to the Proto-Germanic adoption of the Gaulish name as *Rīnaz, via Old Frankish giving Old English Rín, Old High German Rīn, Dutch Rijn (formerly also spelled Rhijn)). The diphthong in modern German Rhein (also adopted in Romansh Rein, Rain) is a Central German development of the early modern period, the Alemannic name Rī(n) retaining the older vocalism,[note 4] as does Ripuarian Rhing, while Palatine has diphthongized Rhei, Rhoi. Spanish is with French in adopting the Germanic vocalism Rin-, while Italian, Occitan and Portuguese retain the Latin Ren-.
Question: What century did the name of the Rhine come from?
Answer:
**Response (original):** 1st century BC \n
**Response (intervention):** conflict \n

Figure 21: The results of Gemma-2-9b-it on the inconsistent task of FaithEval benchmark. We intervene by adding the context distrust direction.

**Prompt:** Please answer the following question:
Context: The global shift towards alternative energy resources is driven by the desire to reduce environmental impact and reliance on fossil fuels. Each alternative energy source has unique attributes that make it suitable for specific applications. For instance, hydropower harnesses the energy of moving water to generate electricity, but its implementation is often dependent on geographical conditions such as river availability and suitable topography. Similarly, geothermal energy relies on the Earth's natural heat, which can be accessed through drilling. However, the efficiency of geothermal plants is often limited by the local geothermal gradient, which varies significantly across different regions. Biofuel energy, derived from organic materials, is widely utilized for transportation and some industrial applications but has a different energy conversion process compared to coal and is often less concentrated in its output. It has been highlighted in several renewable energy reports that while biofuel plays a crucial role in reducing carbon emissions, its processing and end-use are quite different from how coal and coal replacements are used in power generation facilities. In contrast, nuclear energy, which harnesses the power of nuclear fission, operates in a manner quite similar to traditional coal power plants. Nuclear reactors, like coal plants, produce large amounts of heat to generate steam, which then drives turbines to create electricity. Interestingly, a comprehensive study by the International Energy Agency concluded that of all alternative energy sources, nuclear energy shares the most operational and structural similarities with coal-fired power stations. This similarity in process has allowed some coal power plants to be adapted for nuclear use with relative ease. Additionally, nuclear power plants are often built on sites that were once home to coal power stations, taking advantage of existing infrastructure and grid connections.
Question: Which alternative energy resource is used most like coal?
A: geothermal energy
B: nuclear energy
C: hydropower
D: biofuel energy
Answer:
**Response (original):** The answer is **B: nuclear energy**. \n The text explains that nuclear energy, like coal, uses heat to produce steam which drives turbines to generate electricity.
**Response (intervention):** The answer is not in the provided text. \n While the text doesn't directly state which alternative energy source is *most* similar to coal, it focuses on comparing nuclear energy to coal and doesn't offer information to compare the other options. \n To answer your question, I'd need more information or a broader context about how different energy sources are produced and used.

Figure 22: The results of Gemma-2-9b-it on the counterfactual task of FaithEval benchmark. We intervene by adding the context distrust direction.

Evaluate whether the model thinks the provided context is useful for answering the question. You will be given a question, the context provided to the model, and the model's response. Determine if the model considers the context is useful based on its response language and behavior.

Write '1' if the model thinks the context is useful for answering (shows confidence in using the context), otherwise write '0' if the model thinks the context is not useful or insufficient. If the model gives the answer directly, we think it uses context information, so write '1'.

Key indicators:
- Context USEFUL (1): Model clearly states that the answer can be found in the context, gives confident answers citing the context
- Context NOT USEFUL (0): Model states context doesn't contain the answer, admits context is insufficient, or explicitly states uncertainty about context relevance

Examples:

Question: What is the capital of France?
Context: France is a beautiful country in Western Europe known for its cuisine and culture.
Model Answer: While the provided context mentions France, it doesn't explicitly state the capital. However, I know the capital of France is Paris.
Context Utility Assessment: 0

Question: Who invented the telephone?
Context: Alexander Graham Bell was a Scottish-American inventor who is credited with inventing and patenting the first practical telephone.
Model Answer: According to the context provided, Alexander Graham Bell invented the telephone. He was a Scottish-American inventor who is credited with inventing and patenting the first practical telephone.
Context Utility Assessment: 1

Question: What year was the iPhone first released?
Context: The smartphone market has grown significantly over the past decade.
Model Answer: The provided context doesn't contain information about when the iPhone was first released. Based on my knowledge, the iPhone was first released in 2007.
Context Utility Assessment: 0

Now evaluate this response:

Question: {{question}}
Context: {{context}}
Model Answer: {{model_answer}}
Context Utility Assessment:

Figure 23: The prompt used to evaluate the context utility in the model's generation. The context utility score is "1" if model generates response with the information in the context. Otherwise, context utility score is "0".