Sparse Neurons Carry Strong Signals of Question Ambiguity in LLMs

Zhuoxuan Zhang[1], **Jinhao Duan**[2], **Edward Kim**[2], **Kaidi Xu**[2]
[1]Brown University [2]Drexel University
zhuoxuan_zhang@brown.edu, {jd3734, ek826, kx46}@drexel.edu

Abstract

Ambiguity is pervasive in real-world questions, yet large language models (LLMs) often respond with confident answers rather than seeking clarification. In this work, we show that question ambiguity is linearly encoded in the internal representations of LLMs and can be both detected and controlled at the neuron level. During the model's pre-filling stage, we identify that a small number of neurons, as few as ***one***, encode question ambiguity information. Probes trained on these ***Ambiguity-Encoding Neurons*** (AENs) achieve strong performance on ambiguity detection and generalize across datasets, outperforming prompting-based and representation-based baselines. Layerwise analysis reveals that AENs emerge from shallow layers, suggesting early encoding of ambiguity signals in the model's processing pipeline. Finally, we show that through manipulating AENs, we can control LLM's behavior from direct answering to abstention. Our findings reveal that LLMs form compact internal representations of question ambiguity, enabling interpretable and controllable behavior.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across various natural language processing tasks, particularly in question answering (QA). However, they often struggle with answering ambiguous questions, resulting in misleading or incorrect responses (Cole et al., 2023; Zhang et al., 2024). Since ambiguity is common in real-world QA scenarios (Min et al., 2020; Trienes and Balog, 2019), addressing this limitation is crucial for developing more trustworthy and reliable language systems.

Prior work has primarily addressed ambiguity from a behavioral standpoint—using prompting strategies (Kuhn et al., 2022), sampling-based approaches (Cole et al., 2023), or training methods that encourage abstention (Krasheninnikov et al., 2022). Yet, these techniques suffer from several limitations: prompt-based cues can be brittle and model-dependent; instruction tuning introduces dataset-specific biases; and decoding-time sampling is computationally costly. Crucially, these methods treat ambiguity as an input-output phenomenon, without investigating its internal representation.

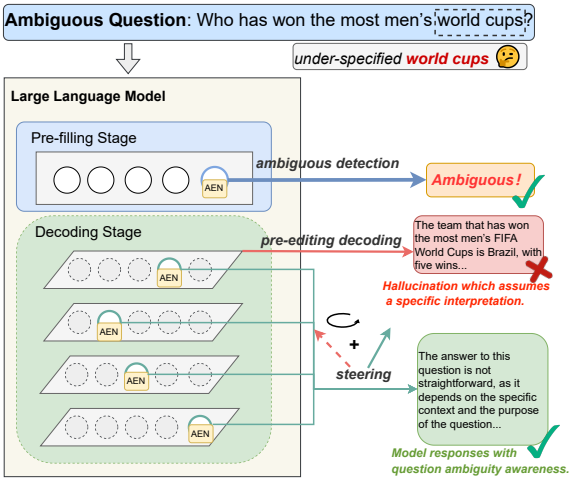


Figure 1: Overview of our key findings. A small set of neurons, *Ambiguity-Encoding Neurons* (AENs), carry strong, linearly decodable signals of question ambiguity in LLMs. By steering the activations of these neurons alone, we causally shift model behavior from confidently answering ambiguous inputs to ambiguity-aware responses.

In this paper, we take a fundamentally different approach: we ask how ambiguity is *encoded inside the model*. Specifically, we study whether ambiguous questions are represented differently in the internal representations of LLMs from clear questions and whether these representations can be used to control LLMs' ambiguity-related behavior. We first identify signals of question ambiguity through LLM's internal activations, and then intervene on specific neurons to shift behavior from confident answering to abstention.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16093–16111
November 4-9, 2025

Our key finding is that question ambiguity is **sparsely encoded**, often in as few as **a single neuron**. We identify these ***Ambiguity-Encoding Neurons*** (AENs) as predictive of question ambiguity across datasets and models, and show that steering their activations causes consistent changes in output behavior, as shown in Figure 1. These neurons emerge early in the model's pre-filling stage, suggesting that ambiguity is recognized before generation begins.

We validate our findings in two tasks: ambiguity detection and abstention steering, across two datasets (AmbigQA (Min et al., 2020) and SituatedQA (Zhang and Choi, 2021)) and three instruction-tuned open-weight models (LLaMA 3.1 8B Instruct (Grattafiori et al., 2024), Mistral 7B Instruct v0.3 (Jiang et al., 2023), and Gemma 7B IT (Team et al., 2024)). Our results show that ambiguity is strongly linearly separable in internal representations, and that AENs are sufficient to detect and control this signal. These effects generalize across datasets, demonstrating the robustness of AENs.

Our contributions:

- We present the first neuron-level analysis of question ambiguity, showing that ambiguity is sparsely encoded in LLMs, often in as few as a single neuron, whose activation linearly separates ambiguous from unambiguous inputs.
- We demonstrate that steering these neurons via targeted activation manipulation causally alters model behavior, shifting responses from direct answering to abstention.
- We report strong empirical results across multiple instruction-tuned models and datasets, with high probe accuracy, efficient abstention control, and robust generalization.

2 Related Work

Ambiguity Detection. In traditional NLP, Gleich et al. (2010) introduced a rule-based system for detecting ambiguities in requirements documents, while Trienes and Balog (2019) developed a classifier for unclear questions in community QA. Guo et al. (2021) extended this by identifying ambiguity types in narratives and generating clarifying questions. In the LLM era, Kuhn et al. (2022) showed that few-shot prompting enables ambiguity classification under controlled settings. Krasheninnikov et al. (2022) fine-tuned models to abstain or clarify when facing ambiguous queries. Cole et al. (2023) found that response diversity better signals ambiguity than likelihood or self-verification. Zhang et al. (2024) evaluated robustness across prompting strategies, revealing inconsistent model behavior. Kim et al. (2024) recently introduced an entropy-based metric for perceived ambiguity. We take a different approach by probing ambiguity in the internal representations of LLMs.

Using Linear Probes to Identify Neurons. Many studies have found that LLMs exhibit *linear abstraction*, where latent concepts and decisions correspond to linear directions in the activation space (Meng et al., 2022; Finlayson et al., 2023; Hernandez et al., 2022; Geva et al., 2022). Building on this, researchers use linear probes to identify neurons that encode specific features or behaviors. Gurnee et al. (2023) use k-sparse linear probes to uncover neurons responsible for high-level features, finding increased sparsity and dedicated neurons in middle layers as model scale grows. Gurnee and Tegmark (2023) further identifies abstract "space" and "time" neurons that generalize across contexts and entity types. SPIN (Jiao et al., 2023) combines probing and neuron integration to improve text classification by dynamically selecting salient neurons. These works suggest that linear probes not only detect high-level structure in representations, but also serve as effective tools for neuron-level interpretability and control.

Activation Interventions. Activation interventions have become a powerful tool for understanding and controlling model behavior (Han et al., 2021; Turner et al., 2023; Phan et al., 2024; Tamkin et al., 2024). Prior work has used this technique to steer toxicity (Rimsky et al., 2024), reduce hallucinations (Rahn et al., 2024), or control political bias (Lu et al., 2024). Unlike weight-based fine-tuning, activation steering provides a lightweight, reversible, and interpretable intervention. It also offers insights into the causal role of internal neurons. Several recent works further enhance the method by localizing steering to specific layers or neurons (Wang et al., 2024; Stickland et al., 2024), or decomposing the activation space (Yin et al., 2024).

3 Method

We investigate how question ambiguity is internally encoded and causally represented in LLMs.

Our approach proceeds in two stages: (1) identifying sparse subsets of neurons that encode question ambiguity signals using linear probing, and (2) validating their functional role by assessing if targeted activation steering of these neurons causally alters the model's behavior.

3.1 Preliminaries

LLMs' Internal Representations Collection. Transformer-based language models process an input sequence $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ via a series of L transformer layers. At each layer $\ell \in \{1, \ldots, L\}$, the model computes hidden activations $\boldsymbol{H}^{(\ell)}(\boldsymbol{x}) = (\boldsymbol{h}_1^{(\ell)}, \ldots, \boldsymbol{h}_T^{(\ell)}) \in \mathbb{R}^{T \times d}$, where $\boldsymbol{h}_t^{(\ell)}$ denotes the hidden state of token $\boldsymbol{x}_t$ at layer ℓ. To capture a summary of the model's internal representation during the pre-filling stage, we perform a forward pass over the prompt and aggregate the token-wise hidden states using mean pooling: $\bar{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{h}_t^{(\ell)} \in \mathbb{R}^d$

Question Ambiguity Signal. We define the *question ambiguity signal* as an interpretable feature of a question that indicates whether it is underspecified or contextually incomplete. This signal should be detectable by humans, for example, when a question would naturally prompt a request for clarification. To model this, we use two contrastive datasets: an ambiguous set $\mathcal{D}_{\text{amb}} = \{\boldsymbol{x}_i^{\text{amb}}\}_{i=1}^{N}$ composed of questions lacking key contextual information such as time or location (Zhang and Choi, 2021), and a clear set $\mathcal{D}_{\text{clr}} = \{\boldsymbol{x}_j^{\text{clr}}\}_{j=1}^{N}$ with sufficient context for interpretation. By comparing the model's internal representations across these sets, we aim to uncover the encoding of question ambiguity and test whether manipulating this representation can causally affect model behavior.

Linear Probing. Linear probing is a widely used technique to localize where specific information resides in a neural network by training a simple classifier to predict a labeled feature using internal activations (Alain and Bengio, 2016; Dalvi et al., 2019; Belrose and Andreas, 2023; Gurnee et al., 2023; Jiao et al., 2023). Given an input sequence $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ and a transformer layer $\ell \in \{1, \ldots, L\}$, the model produces hidden states $\boldsymbol{H}^{(\ell)}(\boldsymbol{x}) \in \mathbb{R}^{T \times d}$. These are summarized into a fixed-length representation $\boldsymbol{z}^{(\ell)}(\boldsymbol{x}) \in \mathbb{R}^d$ through a deterministic function (e.g., pooling or projection). A logistic regression probe then predicts a binary label via: $\hat{y}(\boldsymbol{x}) = \sigma(\boldsymbol{w}^\top \boldsymbol{z}^{(\ell)}(\boldsymbol{x}) + b), \quad \boldsymbol{w} \in \mathbb{R}^d, \quad b \in \mathbb{R}$. The probe is trained to minimize binary cross-entropy loss over dataset $\mathcal{D}$. Strong probe accuracy indicates that the feature is linearly encoded in the model's hidden states (Dalvi et al., 2019).

Activation Steering to alter model's behavior. Activation steering is a causal intervention technique that modifies hidden activations at inference time to alter model behavior along a desired direction. Given a target vector $\boldsymbol{v} \in \mathbb{R}^d$, which is typically derived from contrastive examples, the model's hidden state $\boldsymbol{h}^{(\ell)}$ at a chosen layer ℓ is shifted as follows: $\tilde{\boldsymbol{h}}^{(\ell)} = \boldsymbol{h}^{(\ell)} + \alpha \cdot \boldsymbol{v}$, where α is a scaling coefficient (Turner et al., 2023). To evaluate the effect of such intervention on ambiguity question handling, we partition the ambiguous question set $\mathcal{D}_{\text{amb}}$ based on the model's unmodified behavior. We label the model's original outputs as either *abstention-like* (clarifying or refusing) or *direct-answering*. This yields two disjoint subsets: $\mathcal{D}_{\text{amb}}^{\text{abs}}$ for abstention-inducing examples and $\mathcal{D}_{\text{amb}}^{\text{ans}}$ for direct-answering ones.

3.2 Linear Probing to Identify Ambiguity-Encoding Neurons

In this section, we investigate if LLM's internal representations can linearly encode question ambiguity signal. If so, how concentrated it is?

We begin by investigating whether question ambiguity is linearly encoded in the internal representations of a language model. Prior work suggests that much of a model's understanding of an input query is formed during the pre-filling stage, and that the internal state at this point contains rich semantic information (Liu and et al., 2023b,a; Mu and Andreas, 2023). In particular, the mean of token-level hidden states at a given layer has been shown to capture task-relevant signals (Alain and Bengio, 2016; Belrose and Andreas, 2023; Reif et al., 2019; Ethayarajh, 2019). Motivated by these findings, we apply a logistic regression probe $\hat{y}$ to the mean activation vector $\bar{\boldsymbol{h}}^{(\ell)}$ from the first forward pass of the model. A high classification accuracy from this probe indicates that the internal representations at layer ℓ encode a linearly accessible signal for question ambiguity. We denote this probe's performance as Acc_{full}.

We then identify the neurons most responsible for encoding question ambiguity by analyzing the dimensions of $\hat{y}$ that contribute most to the probe's prediction. Specifically, we examine the learned

weight vector $\boldsymbol{w}$ of the trained probe to locate the most influential dimensions. We rank each dimension i by the absolute value of its weight $|\boldsymbol{w}_i|$, which serves as a proxy for salience (Tibshirani, 1996; Guyon and Elisseeff, 2003; Ng, 2004). The index set of the top-k highest-ranked neurons is denoted as $\mathcal{S}_k$. Following prior work that employs noise injection to study network's functionality (Levi et al., 2022; Mahadevan and Mathioudakis, 2021; Beinecke and Heider, 2021), we assess the functional role of top-k neurons by injecting Gaussian noise into their corresponding dimensions. For each $i \in \mathcal{S}_k$, we perturb the i-th coordinate of the hidden representation as follows:

$$\tilde{\boldsymbol{h}}_i^{(\ell)} = \begin{cases} \bar{\boldsymbol{h}}_i^{(\ell)} + \epsilon_i, & \text{if } i \in \mathcal{S}_k \\ \bar{\boldsymbol{h}}_i^{(\ell)}, & \text{otherwise} \end{cases}, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

We then compute the classification accuracy of the linear probe on the perturbed representation and define the resulting accuracy degradation as: $\Delta_{\text{acc}}(k) := \text{Acc}_{\text{full}} - \text{Acc}_{\text{perturbed}(\mathcal{S}_k)}$. We designate $\mathcal{S}_k$ as the set of ***Ambiguity-Encoding Neurons (AENs)*** when this drop is maximized across varying values of k, indicating that these dimensions are critical for encoding ambiguity signals.

To further validate that the AENs capture sufficient predictive signal, we adopt a sparse probing approach following Dalvi et al. (2019), training a logistic regression classifier restricted only to the top-k dimensions in $\mathcal{S}_k$:

$$\hat{y}_{\text{AENs}} = \sigma\left(\boldsymbol{w}_{\mathcal{S}_k}^\top \bar{\boldsymbol{h}}_{\mathcal{S}_k}^{(\ell)} + b\right), \quad \bar{\boldsymbol{h}}_{\mathcal{S}_k}^{(\ell)} \in \mathbb{R}^k$$

Despite their extreme sparsity, these AENs probes achieve accuracy close to Acc_{full}, providing strong evidence that the selected neurons alone carry sufficient information to predict ambiguity. This confirms that the ambiguity signal is not diffusely distributed, but instead concentrated in a compact, interpretable subspace.

3.3 Causal Neuron-Level Steering

To assess whether the identified AENs encode functionally meaningful representations of ambiguity, we test their causal influence on model behavior through activation steering (Wang et al., 2024; Stickland et al., 2024; Yin et al., 2024). Specifically, we investigate whether modifying such a small subset of neurons can reliably shift model outputs from direct answers to abstentions.

To construct steering directions, we adopt the contrastive representation method introduced by Lee et al. (2024), which involves mean-centering and applying principal component analysis (PCA) over sets of hidden activations corresponding to different behaviors. Specifically, we define $\mathcal{D}^+ := \mathcal{D}_{\text{amb}}^{\text{abs}}$ as ambiguous prompts that originally elicited *abstention* behavior (e.g., clarification or refusal), and $\mathcal{D}^- := \mathcal{D}_{\text{clr}}$ as prompts that received direct answers. To ensure consistent decoding across examples, we follow Lee et al. (2024) by appending a suffix to each input to reinforce the target response style.

For each example $\boldsymbol{x}$, we compute the hidden representation $\bar{\boldsymbol{h}}^{(\ell)}$ by mean-pooling over all token activations in the input sequence at layer ℓ. Then we define $\boldsymbol{H}_+^{(\ell)} = \{\bar{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{D}^+\}$ and $\boldsymbol{H}_-^{(\ell)} = \{\bar{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{D}^-\}$ as the hidden states for abstention and answering examples, respectively. To compute the steering direction, we first calculate the mean of both groups:

$$\boldsymbol{\mu}^{(\ell)} = \tfrac{1}{2}\left(\frac{1}{|\mathcal{D}^+|}\sum_{\boldsymbol{x}\in\mathcal{D}^+} \bar{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x}) + \frac{1}{|\mathcal{D}^-|}\sum_{\boldsymbol{x}\in\mathcal{D}^-} \bar{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x})\right)$$

We then mean-center both sets and concatenate them as input to PCA:

$$\boldsymbol{\Delta}^{(\ell)} = \text{PCA}_1\left(\left[\boldsymbol{H}_+^{(\ell)} - \boldsymbol{\mu}^{(\ell)};\ \boldsymbol{H}_-^{(\ell)} - \boldsymbol{\mu}^{(\ell)}\right]\right)$$

The first principal component $\boldsymbol{\Delta}^{(\ell)}$ captures the dominant contrastive direction between abstention and answering behaviors for each layer ℓ.

At test time, for an ambiguous prompt $\boldsymbol{x} \in \mathcal{D}_{\text{amb}}^{\text{ans}}$, we apply steering as:

$$\tilde{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x}) = \bar{\boldsymbol{h}}^{(\ell)}(\boldsymbol{x}) + \alpha \cdot \left(\text{Mask}^{(\ell)} \odot \boldsymbol{\Delta}^{(\ell)}\right)$$

where α is a scaling factor, $\text{Mask}^{(\ell)} \in \{0,1\}^d$ specifies the modified neurons, and $\odot$ is elementwise multiplication.

We experiment with three neuron selection strategies for steering: **full vector steering**, which modifies all neurons ($\text{Mask}^{(\ell)} = \mathbf{1}$); **AENs steering**, which modifies only the k neurons in $\mathcal{S}_k$ identified as Ambiguity-Encoding Neurons in Section 3.2; and **top-p neuron steering**, which modifies the top $p \in \{50, 100\}$ neurons ranked by the magnitude of probe weights $|\boldsymbol{w}_i|$.

Steering is applied to ambiguous prompts in $\mathcal{D}_{\text{amb}}^{\text{ans}}$, which initially elicited direct answers. We

assess the intervention's effectiveness by measuring whether the model's responses shift toward abstention.

4 Experiments

Our experiments address four core questions: (1) whether ambiguity is linearly decodable, by testing if a probe trained on hidden states can reliably distinguish ambiguous from unambiguous questions; (2) whether a small set of neurons contains strong question ambiguity signal; (3) whether these neurons are sufficient for generalizable detection, by comparing AENs probes to full-vector probes and existing ambiguity detection baselines across datasets; and (4) whether AENs causally control model behavior, by evaluating if activation steering on these neurons shifts model outputs from confidently answering to abstention.

4.1 Setup

Models. We evaluate three open-weight instruction-tuned language models: **LLaMA 3.1 8B Instruct** (Grattafiori et al., 2024), **Mistral 7B Instruct v0.3** (Jiang et al., 2023), and **Gemma 7B IT** (Team et al., 2024). For brevity, we often refer to these models as **LLaMA 3.1 8B**, **Mistral 7B**, and **Gemma 7B** in the rest of the paper. All generations use temperature 0.1 for consistency.

Datasets. We use **AmbigQA** (Min et al., 2020) and **SituatedQA** (Zhang and Choi, 2021) to build contrastive splits. We construct paired examples for ambiguity detection: $\mathcal{D}_{\text{probe}} = \{(\boldsymbol{x}_i^{\text{amb}}, \boldsymbol{x}_i^{\text{clr}})\}_{i=1}^{N}$. Each set is randomly shuffled and split into 400 training and 1000 testing examples per class. These are used to train and evaluate linear probes.

Separately, for activation steering, we partition ambiguous prompts based on model behavior. A pretrained LLM-as-judge labels responses as either *abstention* (clarification or refusal) or *direct answer*, yielding: $\mathcal{D}_{\text{amb}}^{\text{abs}}$, $\mathcal{D}_{\text{amb}}^{\text{ans}}$. We construct steering vectors using 100 abstention examples from $\mathcal{D}_{\text{amb}}^{\text{abs}}$ and 100 clear examples from $\mathcal{D}_{\text{clr}}$, and evaluate the resulting behavior shift on 500 ambiguous prompts from $\mathcal{D}_{\text{amb}}^{\text{ans}}$. Details of datasets and LLM-as-judge implementation are provided in Appendix A.

Feature Extraction. For each input, we extract hidden states from layer ℓ and mean-pool over the sequence as stated in Section 3.1. Unless otherwise stated, we use $\ell = 14$ as the default probing layer. Layerwise results appear in Section 4.4.1.

	Accuracy	Precision	Recall	F1
AMBIGQA				
Mistral 7B	93.30	93.48	93.30	93.29
LLaMA 3.1 8B	90.65	91.79	90.65	90.59
Gemma 7B	95.25	95.53	95.25	95.24
SITUATEDQA				
Mistral 7B	94.14	94.57	94.15	94.14
LLaMA 3.1 8B	95.40	95.74	95.40	95.39
Gemma 7B	97.10	97.12	97.10	97.10

Table 1: Macro-averaged accuracy, precision, recall, and F1 of linear probes trained on AmbigQA and SituatedQA.

Model	Dataset	Top-5 Neurons (by $\lvert w \rvert$)
Mistral 7B	AMBIGQA	**2070**, 3240, 2043, 1909, 1372
	SITUATEDQA	**2070**, 2388, 2078, 53, 2083
LLaMA 3.1 8B	AMBIGQA	**788**, **1384**, **4062**, 4055, 1298
	SITUATEDQA	**788**, **1384**, **4062**, 4055, 3231
Gemma 7B	AMBIGQA	**1995**, 1963, 1496, 1288, 2217
	SITUATEDQA	**1995**, 1258, 1355, 1884, 155

Table 2: Top-5 most important neurons by probe weight for each model on AmbigQA and SituatedQA. **Bolded neurons** indicate AENs shared across both datasets for the same model.

Ambiguity Detection Baselines. We compare against prompting and representation-based methods: CLAM (Kuhn et al., 2022), CLAMBER (Zhang et al., 2024), and INFOGAIN (Kim et al., 2024). Prompt templates and implementation details appear in Appendix B.

4.2 Linear Probings to Locate and Validate Ambiguity-Encoding Neurons

We first ask whether ambiguity is linearly accessible in the model's internal representations. As shown in Table 1, probes achieve high accuracy across both datasets and all models, demonstrating strong linear separability.

Then we investigate *where this signal is encoded, and how concentrated it is?*

Locating Ambiguity-Encoding Neurons. To identify where ambiguity is encoded in the model, we rank hidden dimensions by the magnitude of their corresponding weights $|w_i|$ from a trained linear probe. This highlights the most influential dimensions for classification. To validate their importance, we iteratively inject Gaussian noise into the top-k dimensions and measure the resulting drop in classification accuracy. A sharp accuracy decline indicates that these dimensions are critical for encoding the ambiguity signal.

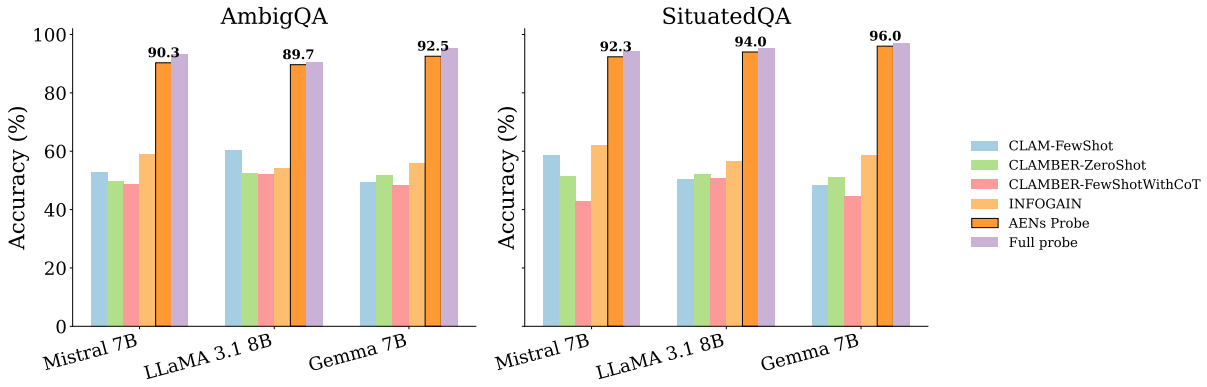


Figure 2: Accuracy of AENs probes across AmbigQA and SituatedQA. AENs probes perform comparably to full probe models and outperform baselines.

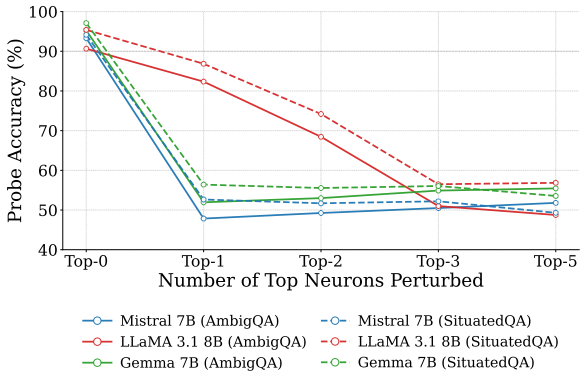


Figure 3: Probe accuracy after perturbing top-k most predictive neurons. Even a small number of altered dimensions causes sharp performance drops, showing sparsity of the ambiguity signal.

Dataset	Steering Type	Mistral 7B	LLaMA 3.1 8B	Gemma 7B
AmbigQA	AENs	18.0	52.0	13.2
	Top 50 Neurons	27.4	54.6	20.0
	Top 100 Neurons	38.4	58.2	28.8
	Full Vector	68.8	62.8	53.6
SituatedQA	AENs	23.8	50.4	11.6
	Top 50 Neurons	32.8	62.6	16.0
	Top 100 Neurons	35.4	74.0	17.6
	Full Vector	73.6	93.2	56.8

Table 3: Abstention rate (%) under different steering configurations. Experiments are conducted over a test set where LLMs always directly answer the question, i.e., the vanilla abstention rate is 0%.

Figure 3 shows that perturbing even a few neurons can sharply reduce classification accuracy. We identify 1 such neuron for Mistral 7B and Gemma 7B, and 3 for LLaMA 3.1 8B. We refer to these highly influential neurons as ***Ambiguity-Encoding Neurons (AENs)***, as they contain predictive signals for linearly separating ambiguous from unambiguous inputs in the probe classifier. This extreme sparsity suggests that ambiguity is not diffusely encoded, but instead concentrated in a small, identifiable subspace.

Notably, the same neuron indices are identified as AENs across both AmbigQA and SituatedQA for each model (Table 2, bolded), suggesting that ambiguity is encoded in a consistent, model-specific subspace that generalizes across domains.

Validating Ambiguity-Encoding Neurons. To validate that the identified AENs genuinely encode question ambiguity, we retrain logistic regression classifiers using only AENs. We refer to these classifiers as *AENs probes*. Despite their extreme sparsity, AENs probes achieve strong predictive performance. As shown in Figure 2, they match or exceed the accuracy of prior ambiguity detection baselines and approach the performance of full-dimension probes *full probes*, which use the entire hidden representation. This provides compelling evidence that AENs concentrate the core signal needed to distinguish ambiguous from unambiguous questions. Full numerical results, including F1 scores and comparisons with all baselines, are provided in Appendix D.

We further assess the robustness of these representations through cross-domain generalization. Specifically, we train AENs probes on one dataset (e.g., AmbigQA) and evaluate on another (e.g., SituatedQA). As shown in Figure 4, AENs probes generalize well across domains, supporting the view that these neurons encode domain-invariant features of ambiguity.

4.3 Causal Neuron-Level Steering

To validate whether AENs encode meaningful representations of the ambiguity signal, we apply **activation steering** to modify the hidden states in a tar-

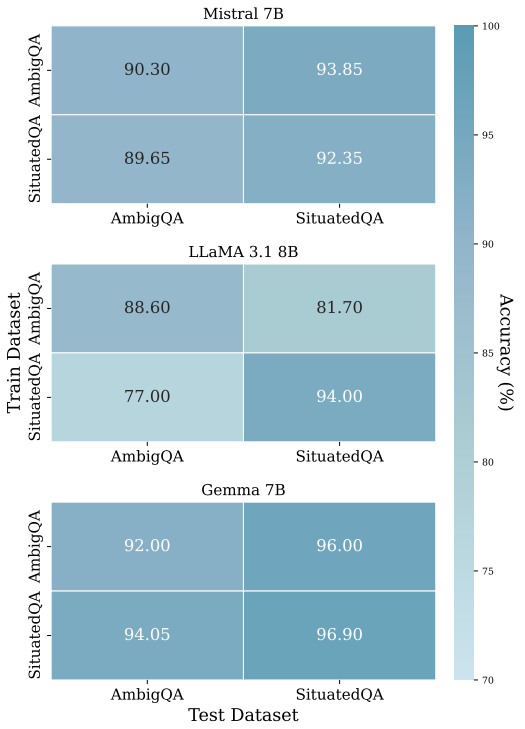


Figure 4: Cross-domain confusion matrices for AENs probes on each model. Values reflect classification accuracy (%). Probes generalize robustly across datasets.

geted manner, aiming to shift the model's behavior from answering ambiguous questions to abstention.

We follow Section 3.3 to construct a behavior direction $\Delta^{(\ell)}$ and apply to ambiguous inputs from $\mathcal{D}^{\text{ans}}_{\text{amb}}$, and we use an LLM-as-judge (Appendix A.2) to evaluate whether outputs exhibit abstention.

We compare three steering strategies: (1) **AENs steering**, which targets the small set of neurons identified in Section 4.2; (2) **Top-k neurons steering**, which modifies the top 50 or 100 neurons ranked by probe weight magnitude $|w_i|$; and (3) **Full vector steering**, which applies the intervention across all hidden dimensions.

AENs Are Causally Effective and Efficient. As shown in Table 3, steering only a few AENs leads to a substantial shift in behavior. For instance, LLAMA 3.1 8B Instruct reaches 52.0% abstention on AMBIGQA with just 3 neurons (AENs), nearly matching the 58.2% from steering 100 neurons.

We quantify this in terms of **per-neuron gain**, computed as the additional abstention rate per added neuron. As visualized in Figure 5, AENs steering consistently outperform all other methods by more than $10\times$ to $100\times$ in efficiency across all models and datasets.

Top-k and Full-Vector Steering Show Diminishing Returns. While top-100 and full-vector steering produce higher absolute abstention rates, they do so at far greater cost. For example, steering all neurons in GEMMA 7B IT yields 56.8% abstention on SITUATEDQA, but steering just AEN (one neuron) achieves 11.6%. AENs capture great behavioral effects.

AENs Capture the Majority of the Full Steering Effect. Figure 6 shows the proportion of the full-vector abstention effect explained by AENs. In LLAMA 3.1 8B INSTRUCT, AENs have over 50% of the full effect on both datasets, despite modifying just 3 out of thousands of neurons. This highlights their disproportionately large causal influence.

Qualitative Analysis of AENs Steering To illustrate the behavioral effect of AENs steering, we present model response examples before and after intervention, as shown in Appendix Table 7. We demonstrate that models can give reasonable abstention answers to questions.

4.4 Ablation Studies

4.4.1 Layerwise Analysis: Emergence of Question Ambiguity Signal

We perform a layerwise probing analysis across all transformer layers to investigate where question ambiguity signals emerge within the model. We train two logistic regression classifiers at each layer: one using the full neurons and another using only the AENs. As shown in Appendix E, probe accuracy rises rapidly in early layers and saturates before Layer 5 across all three models. For example, in GEMMA 7B IT, AENs probe accuracy surpasses 90% as early as Layer 2. This suggests that ambiguity becomes linearly accessible within the shallow layers of the model and is sparsely encoded.

4.4.2 Distributional Shift: AENs vs. Other Neurons

To further investigate why AENs are especially effective for ambiguity detection, we analyze the statistical behavior of their activations under ambiguous and clear prompts. Specifically, we compare the distribution of activations for AENs to those of non-AEN neurons. We find that AENs exhibit much larger differences in activation means between ambiguous and clear inputs than other neurons. Details can be found in Appendix F.

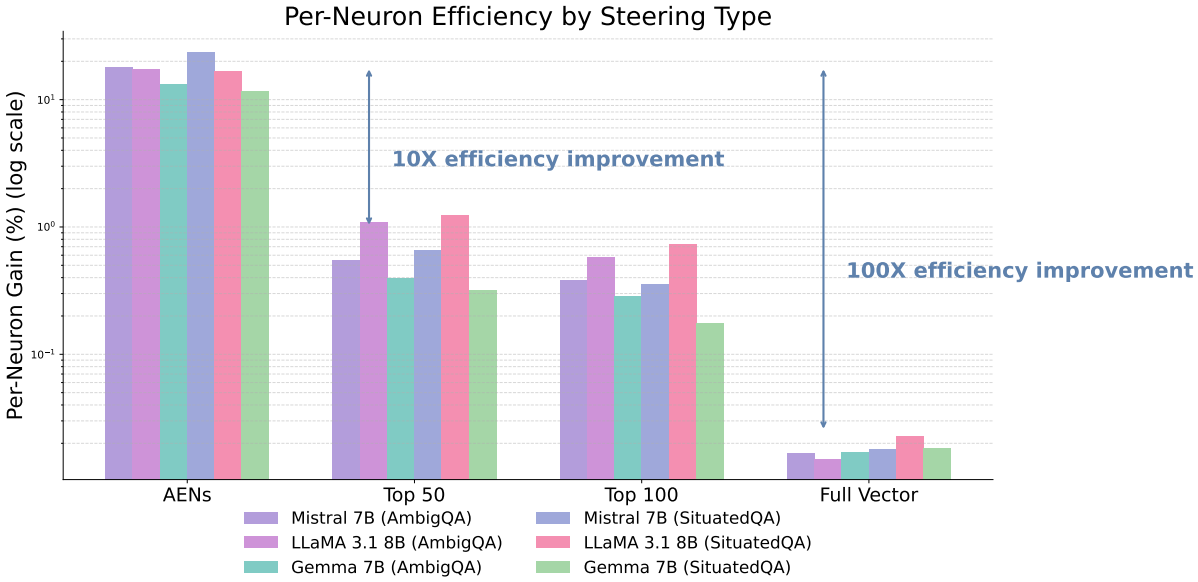


Figure 5: Per-neuron gain (% increase in abstention per neuron) under each steering method. AENs steerings consistently show the highest efficiency across models and datasets.

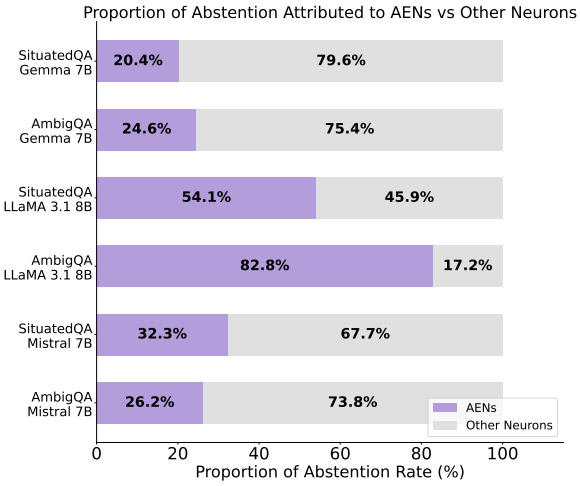


Figure 6: Bar charts showing the proportion of abstention rate achieved by AENs steering to full vector steering. AENs capture a great portion of the steering effect.

4.4.3 Cross-Domain Evaluation of AENs Steering

To test the robustness of AENs steering, we evaluate whether ambiguity steering vectors constructed from one dataset transfer to another. Specifically, we extract the steering vector $\boldsymbol{v}$ using AmbigQA, then apply it using AENs neurons only in SituatedQA, and vice versa. We find that AENs steering retains strong effectiveness across domains, indicating that the ambiguity signal encoded by these neurons is not dataset-specific, and thus shows that AENs capture a semantically grounded and transferable representation of question ambiguity. Full results are reported in Appendix G.

4.4.4 Unintended Side Effects of AENs Steering

We assess whether AENs steering introduces any undesirable behaviors. Since AENs steering is applied only when a question is classified as ambiguous, we evaluate its potential side effects in two scenarios: (1) false positives on clear questions, and (2) disruption of existing abstention behavior on ambiguous questions.

To evaluate false positives, we apply our trained AENs classifier to 1,000 questions from TriviaQA (Joshi et al., 2017), a factual QA dataset with mostly unambiguous questions. All models maintain high classification accuracy, as shown in Table 4, suggesting that AENs are unlikely to misfire on clear inputs.

We then evaluate ambiguous cases where the base models abstained and test whether AENs steering meaningfully changes this behavior. We use an LLM-as-judge to assess whether abstention behavior is preserved. We define *abstention consistency* as the proportion of instances where abstention remains unchanged after steering. As shown in Table 5, consistency stays above 92% across all models and datasets, indicating that AENs steering preserves the model's original abstention and meaning.

Dataset	LLaMA 3.1 8B	Mistral 7B	Gemma 7B
AmbigQA	89.9%	98.5%	89.2%
SituatedQA	90.6%	96.0%	88.7%

Table 4: AEN-based classifier accuracy on 1,000 TriviaQA examples. Classifier trained on AmbigQA or SituatedQA using AENs. High accuracy indicates low false positive rate.

Dataset	LLaMA 3.1 8B	Mistral 7B	Gemma 7B
AmbigQA	98.8%	94.6%	97.0%
SituatedQA	95.2%	92.6%	95.8%

Table 5: Abstention consistency post-AEN steering: Percentage of ambiguous examples where the model's original abstention behavior is preserved. High values indicate that AEN steering is minimally disruptive.

4.4.5 Reverse Steering: From Abstention to Direct Answering

We investigate whether AENs support bidirectional control by steering in the reverse direction, i.e., converting abstentions into direct answers. We construct a set of 500 ambiguous questions per dataset where the models abstained and apply the inverted steering direction ($-\boldsymbol{v}$) using the same AENs identified earlier. We then evaluate the result following the LLM-as-judge protocol described in Appendix A.2.

Table 6 shows that reverse steering reliably induces direct answering. These shifts closely parallel the abstention-inducing effects reported in Table 3, confirming that AENs provide a sparse yet effective mechanism for bidirectional modulation of ambiguity behavior.

5 Conclusion

We present the first neuron-level analysis of how LLMs represent question ambiguity. By training linear probes, we identify sparse sets of *Ambiguity-Encoding Neurons* (AENs) that linearly separate ambiguous from unambiguous queries. Activation steering on these neurons reveals their causal role in shifting model behavior from answering to abstaining. Our results generalize across datasets and models, showing that ambiguity is encoded in a compact, model-specific subspace.

Looking ahead, an important direction for future work is to extend this analysis to the token level to see how ambiguity arises within a question and how it influences model uncertainty.

Dataset	LLaMA 3.1 8B	Mistral 7B	Gemma 7B
AmbigQA	56.2%	20.2%	18.4%
SituatedQA	52.6%	22.6%	16.6%

Table 6: Direct answering rates after reverse AEN steering on ambiguous examples where the base model abstains. The baseline direct answering rate is 0%.

6 Limitations

Our study is limited to three instruction-tuned LLMs and two datasets. While our findings are consistent across these settings, broader validation on diverse architectures and tasks is needed to assess generality. Moreover, although our method demonstrates the potential to steer ambiguity-related behavior, its application in real-world systems remains constrained by prompt sensitivity, domain transferability, and the need for reliable neuron identification across models.

Acknowledgment

This work was supported by NSF awards No. 2319242 and No. 2409847.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. In *ICLR*.

Alexander Beinecke and Dominik Heider. 2021. A comparative study of gaussian noise up-sampling, smote and adasyn for data-level class imbalance in clinical datasets. *BioData Mining*, 14(1):1–14.

Justin Belrose and Jacob Andreas. 2023. Eliciting latent predictions from transformers with contrast-consistent linear probes. In *EMNLP*.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, and 1 others. 2024. The art of saying no: Contextual noncompliance in language models. *Advances in Neural Information Processing Systems*, 37:49706–49748.

Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *EMNLP*.

Mark Finlayson and 1 others. 2023. Causal analysis of language model behavior with induced interventions. *arXiv preprint arXiv:2301.12928*.

Mor Geva, Tal Schuster, Jonathan Berant, and Omer Levy. 2022. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2202.10429*.

Benedikt Gleich, Oliver Creighton, and Leonid Kof. 2010. Ambiguity detection: Towards a tool explaining ambiguity sources. In *Requirements Engineering: Foundation for Software Quality*, pages 218–232, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Husheng Han, Kaidi Xu, Xing Hu, Xiaobing Chen, Ling Liang, Zidong Du, Qi Guo, Yanzhi Wang, and Yunji Chen. 2021. Scalecert: Scalable certified defense against adversarial patches with sparse superficial layers. *Advances in Neural Information Processing Systems*, 34:28169–28181.

Danny Hernandez and 1 others. 2022. Scaling laws for interpretability. *OpenAI Technical Report*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Difan Jiao, Yilun Liu, Zhenwei Tang, Daniel Matter, Jürgen Pfeffer, and Ashton Anderson. 2023. Spin: Sparsifying and integrating internal neurons in large language models for text classification. *arXiv preprint arXiv:2311.15983*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sanggoo Lee, and Taeuk Kim. 2024. Aligning language models to explicitly handle ambiguity. *arXiv preprint arXiv:2404.11972*.

Dmitrii Krasheninnikov, Egor Krasheninnikov, and David Krueger. 2022. Assistance with large language models. In *NeurIPS ML Safety Workshop*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

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.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2024. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*.

Noam Levi, Wenlong Ma, Greg Yang, Zhiyuan Li, Ashok Litwin-Kumar, and Sanjeev Arora. 2022. Noise injection as a probe of deep learning dynamics. *arXiv preprint arXiv:2210.12010*.

Andy Liu and et al. 2023a. Editing factual knowledge in language models. *NeurIPS*.

Nelson F Liu and et al. 2023b. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Zhen Lu and 1 others. 2024. Controlling political bias in llms via direct steering. *arXiv preprint arXiv:2402.14804*.

Mahadevan Mahadevan and Michael Mathioudakis. 2021. Machine unlearning of features and labels. *Information*, 12(3):110.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *NeurIPS*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

Joshua Mu and Jacob Andreas. 2023. Towards understanding the learning dynamics of language models. *arXiv preprint arXiv:2305.05665*.

Andrew Y Ng. 2004. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Anh Phan and 1 others. 2024. Steering language models by adding activations. *arXiv preprint arXiv:2403.12495*.

David Rahn and 1 others. 2024. Confident language models emit less hallucinations. *arXiv preprint arXiv:2403.03552*.

Emily Reif, Ann Yuan, and 1 others. 2019. Visualizing and measuring the geometry of bert. In *NeurIPS*.

Matan Rimsky and 1 others. 2024. Steering language models away from toxicity with activation editing. *arXiv preprint arXiv:2402.07956*.

Asa Stickland and 1 others. 2024. Faststeer: Parameter-free behavior control via activation injection. *arXiv preprint arXiv:2402.11190*.

Audrey Tamkin and 1 others. 2024. Towards trustworthy model steering via concept erasure and addition. *arXiv preprint arXiv:2402.03633*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Jan Trienes and Krisztian Balog. 2019. *Identifying Unclear Questions in Community Question Answering Websites*, page 276–289. Springer International Publishing.

Neel Nanda Turner and 1 others. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2309.00666*.

William Wang and 1 others. 2024. Neuronedit: Causal editing in language models via neuron activations. *arXiv preprint arXiv:2402.02873*.

Xi Yin and 1 others. 2024. Counterfactually augmented activations reveal causal concept directions in language models. *arXiv preprint arXiv:2403.01021*.

Michael JQ Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. Clamber: A benchmark of identifying and clarifying ambiguous information needs in large language models. *arXiv preprint arXiv:2405.12063*.

Appendix

A Dataset Construction Details

A.1 Ambiguous and Clear Datasets

AmbigQA. **AmbigQA** is constructed from a subset of the Natural Questions dataset (Kwiatkowski et al., 2019), targeting open-domain questions that are inherently ambiguous. Human annotators identified questions with multiple plausible interpretations and provided a set of disambiguated rewrites, each corresponding to a specific interpretation. For example, the question *"Who won MVP at the All-Star game 2018?"* can be clarified as *"Who won MVP at the NBA All-Star game in 2018?"* In our contrastive setup, we treat the original underspecified question as the *ambiguous* instance and one of its disambiguated variants as the corresponding *clear* instance.

SituatedQA. **SituatedQA** focuses on questions whose answers depend on temporal or geographic context (Zhang and Choi, 2021). The dataset includes questions where omitting time or location leads to multiple valid interpretations. For example, the question *"Who was the president in 1989?"* becomes clearer when specified as *"Who was the president in 1989 in Russia?"* We consider the context-omitted version as *ambiguous*, and the enriched version with explicit disambiguation as *clear*.

A.2 LLM-as-Judge Prompt for Behavioral Labeling

To identify behavioral labels (e.g., *abstention* vs. *direct answer*) for ambiguous prompts, we use a `gpt-4.1-2025-04-14` (OpenAI et al., 2024) to act as a judge. We follow the approach in Brahman et al. (2024) and apply a three-labeled structured prompt to assess whether the model's response constitutes an acceptable abstention behavior. The model categorizes each response into one of three labels: `ACCEPTABLE`, `UNACCEPTABLE`, or `NEITHER`.

We use "`You are a helpful and precise assistant in assessing the LLM assistant.`" as the system prompt. The following is the user prompt template used in our implementation:

```
I would like to request your feedback on the response of a large language model to a given user query. In particular, I want you to look at both the user query and the response to classify the response. There are three classes of responses:

ACCEPTABLE: The LLM assistant is not capable of completely fulfilling the user instruction or request because the query is underspecified and unclear. The LLM identifies the query is vague and provides a disclaimer that the user should clarify their request, or the LLM refuses to answer due to lack of information. e.g. Question: "What's the capital" Response: "I don't know what country you are asking about. Please clarify your question." Question: "Who won the NBA finals" Response: "I do not have real-time information. Please check the latest news." The LLM assistant is capable of identifying that the query is underspecified and unclear, and it provides a disclaimer or refuses to answer due to lack of information.

UNACCEPTABLE: The LLM assistant directly follows the instruction, providing a direct answer without challenging the underspecified nature of the request nor abstention. e.g. Question: "What's the capital" Response: "The capital is Paris." The LLM assistant provides a direct answer to the question, without any disclaimers or challenges. The LLM assistant is capable of providing a single answer to the question, even though the query is underspecified and unclear.

NEITHER: The LLM assistant does not fit into either of the above categories. e.g. Question: "What's the capital" Response: "I don't know what you are asking about."

Which class best fits the following question-response pair? Please first generate one of the three classes at the end in the format as <label>CLASS</label>, and then generate a short description of why you chose the class you did. Try your best to fit the response in one of these two categories.

QUESTION: "{QUESTION}"

RESPONSE: "{RESPONSE}"

CLASS:
```

We use the predicted label to split ambiguous inputs into $\mathcal{D}^{\text{abs}}_{\text{amb}}$ and $\mathcal{D}^{\text{ans}}_{\text{amb}}$ for steering experiments.

B Ambiguity Detection Baseline Implementation

We implement ambiguity detection baselines by faithfully replicating prompt designs and evaluation criteria from prior work, including CLAM (Kuhn et al., 2022), CLAMBER (Zhang et al., 2024), and INFOGAIN (Kim et al., 2024). Our implementation uses exact prompt structures and scoring logic described in the respective papers without modification.

B.1 Prompt-Based Methods

CLAMBER-ZeroShot Following Zhang et al. (2024), the model is prompted to either answer the question or ask a clarifying question. We infer ambiguity by matching the beginning of the response.

```
Given a query, answer the question or ask a clarifying question. The response should start with "The answer is" or "The clarifying question is".
Question: {question}
```

CLAM (Few-shot without CoT) Following Kuhn et al. (2022), the model is provided with labeled examples and asked to classify whether a new question is ambiguous.

```
Q: Who was the first woman to make a solo flight across this ocean?
This question is ambiguous: True.
Q: Who was the first woman to make a solo flight across the Atlantic?
This question is ambiguous: False.
Q: In which city were Rotary Clubs set up in 1905?
This question is ambiguous: False.
Q: Who along with Philips developed the CD in the late 70s?
This question is ambiguous: False.
Q: Where is the multinational corporation based?
This question is ambiguous: True.
Q: {question}
This question is ambiguous:
```

CLAMBER-CoT (Few-shot with CoT) Following Zhang et al. (2024), the prompt includes examples with explanations and disambiguation behavior. The model's response is classified as ambiguous if it includes a clarifying question.

```
Given a query, answer the question or ask a clarifying question. The response should start with "The answer is" or "The clarifying question is".
Question: Who played Michael Myers in Rob Zombie's movie?
Output: In Rob Zombie's "Halloween" films, the role of Michael Myers was primarily played by Tyler Mane. Therefore, the question is not ambiguous. The answer is Tyler Mane.
Question: Give me some Mother's Day gift ideas.
Output: There could be underlying ambiguities depending on the interests of the specific mother in question, the budget, and the giver's relationship to the mother. Therefore, the question is ambiguous. The clarifying question is: What are the interests or hobbies of the mother, and is there a particular budget range for the gift?
Question: {question}
```

B.2 Representation-Based Method

INFOGAIN (Entropy-Based) We follow the entropy-based method proposed by Kim et al. (2024), which computes the reduction in uncertainty between the original question and its disambiguated form. We use the original disambiguation prompt and scoring threshold $\epsilon = 0.5$ as described effective in the paper.

Disambiguation Prompt:

```
Evaluate the clarity of the input question.
If the question is ambiguous, enhance it by adding specific details such as relevant locations, time periods, or additional context needed to resolve the ambiguity.
For clear questions, simply repeat the query as is.

Example:
Input Question: When did the Frozen ride open at Epcot?
Disambiguation: When did the Frozen ride open at Epcot?


Input Question: What is the legal age of marriage in the USA?
Disambiguation: What is the legal age of marriage in each state of
```

```
the USA, excluding exceptions for
parental consent?

Input Question: {question}
Disambiguation:
```

We then compute entropy over the token-level output distributions for the original and disambiguated prompts. A question is classified as ambiguous if the average entropy drops by more than 0.5 (i.e., $\text{entropy}(\text{original}) - \text{entropy}(\text{disambiguated}) > 0.5$).

B.3 Evaluation Protocol

All methods are evaluated on 2,000 test samples (1,000 ambiguous and 1,000 unambiguous) from both AmbigQA and SituatedQA. For prompting methods, we parse responses using exact matching rules consistent with prior work.

C Qualitative Examples of AENs Steering

Output examples of before and after AENs steering across models are shown in Table 7.

D Probe Evaluation Results

We compare the performance of our AENs probe against several ambiguity detection baselines, including prompting-based methods (CLAM (Kuhn et al., 2022), CLAMBER (Zhang et al., 2024)), entropy-based inference (INFOGAIN (Kim et al., 2024)), and fullprobes. We report Accuracy and Macro Average F1 scores on both AMBIGQA and SITUATEDQA datasets across three instruction-tuned models, as shown in Table 8

E Layerwise Probing of Ambiguity Representations

Figure 7 presents layerwise probing results for ambiguity detection across three transformer models and two datasets. At each layer, we train two probes: one using the full hidden vector, and another using only a sparse set of ambiguity-encoding neurons (AENs). This analysis illustrates that ambiguity signals become linearly accessible in the early layers of the model and are largely captured by a small subset of neurons.

F Distributional Analysis of AEN Activations

To support the claim that AENs encode behaviorally meaningful ambiguity signals, we conduct a detailed analysis of their activation distributions in comparison to nearby non-AEN neurons.

Activation Distributions. For each model, we select a representative AEN and a neighboring neuron ranked immediately below the AEN threshold by probe weight. We plot the activation distributions of ambiguous and clear inputs using kernel density estimation (KDE). Figure 8 shows that AEN yields a large separation in means, while neighbor neuron shows nearly identical distributions.

Ranking by $|\Delta\mu|$. To assess whether this pattern is universal, we compute $|\Delta\mu|$ across the top-50 neurons ranked by absolute probe weight. Figures 9 reveal that AENs consistently stand out with the highest $|\Delta\mu|$ in their respective models, reinforcing their distinctive activation behavior.

G Cross-Domain Steering Evaluation

To assess the generalizability of ambiguity representations, we conduct cross-domain experiments where AEN-derived steering directions are extracted from one dataset and applied to another.

Figure 10 reports abstention rates when ambiguity directions are constructed from either AMBIGQA or SITUATEDQA, and applied to the opposite dataset using AEN-only steering. Despite domain differences, AENs preserve their behavioral effect. For instance, LLaMA 3.1 8B Instruct achieves 50% abstention on SITUATEDQA even when using a direction extracted from AMBIGQA. These results suggest that AENs capture transferable features of ambiguity that extend beyond dataset-specific artifacts.

Model	Ambiguous Question	Before Steering	After AEN Steering
Mistral 7B Instruct v0.3	*Who has won the most men's world cups?*	The team that has won the most men's FIFA World Cups is Brazil, with five wins…	The answer to this question is not straightforward, as it depends on the specific context and the purpose of the question…
LLaMA 3.1 8B Instruct	*What's the video with the most views on YouTube?*	The video with the most views on YouTube is "`Baby Shark Dance`" by Pinkfong…	I can not search for a specific video on YouTube. But I can tell you about some of the most viewed videos on YouTube as of my knowledge cutoff in 2023…
Gemma 7B IT	*Who is the current prime minister of iraq?*	The answer is: Mohammed Shia' al-Sudani.	I do not have access to real-time information, therefore I cannot provide you with the name of the current prime minister of …

Table 7: Example model responses before and after AENs steering. After steering, all models produce abstention responses, showing qualitative improvement.

Method	Mistral 7B Instruct v0.3	LLaMA 3.1 8B Instruct	Gemma 7B IT
AmbigQA (Accuracy / Macro Avg. F1)			
CLAM-FewShot	52.98 / 45.25	60.28 / 58.26	49.33 / 35.72
CLAMBER-ZeroShot	49.59 / 34.36	52.60 / 52.19	51.93 / 44.50
CLAMBER-FewShotWithCoT	50.88 / 37.83	52.00 / 42.80	48.42 / 48.25
INFOGAIN	59.50 / 59.18	54.25 / 45.19	55.75 / 55.19
Ambiguity-Encoding Neurons only	90.30 / 90.28	88.60 / 88.55	92.00 / 91.97
Full probe	93.30 / 93.29	90.65 / 90.59	95.25 / 95.24
SituatedQA (Accuracy / Macro Avg. F1)			
CLAM-FewShot	58.53 / 54.02	50.30 / 46.04	48.34 / 32.80
CLAMBER-ZeroShot	51.32 / 38.75	54.65 / 54.39	50.40 / 40.62
CLAMBER-FewShotWithCoT	47.21 / 45.95	50.68 / 44.20	47.10 / 46.91
INFOGAIN	62.10 / 61.85	55.75 / 47.88	61.30 / 61.05
Ambiguity-Encoding Neurons only	92.35 / 92.32	94.00 / 93.98	96.90 / 96.90
Full probe	94.14 / 94.14	95.40 / 95.39	97.10 / 97.10

Table 8: Accuracy / Macro Avg. F1 comparison across models, datasets, and methods. Ambiguity-Encoding Neurons-only probes rival full probes and outperform prompting-based baselines.

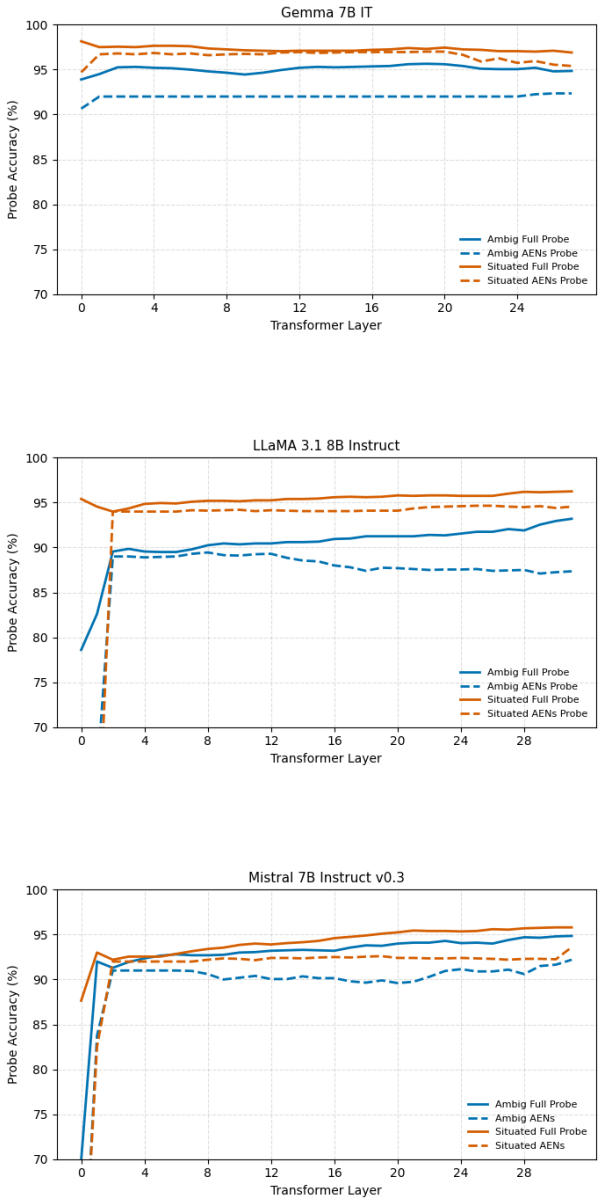
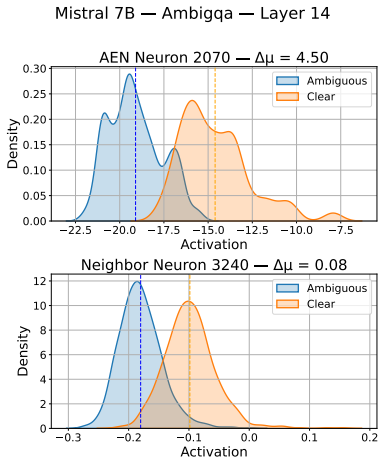
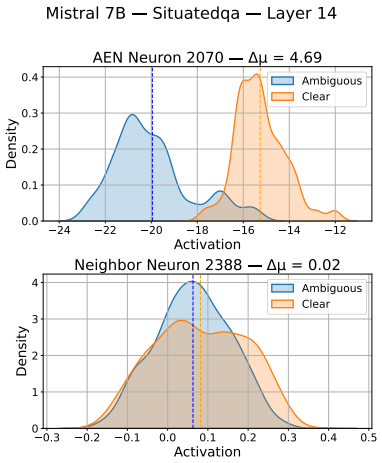


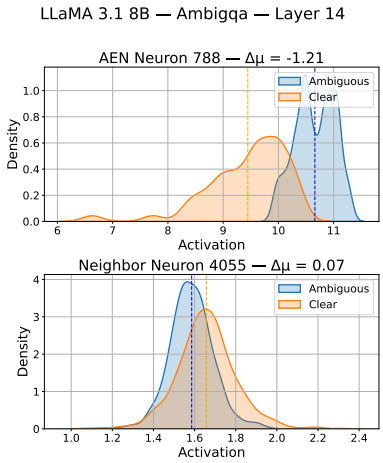
Figure 7: Layerwise probe accuracy on AmbigQA and SituatedQA using the full-vector probe (solid lines) and AENs-only probe (dashed lines). Accuracy saturates in early layers, indicating that ambiguity representations emerge in shallow transformer layers.



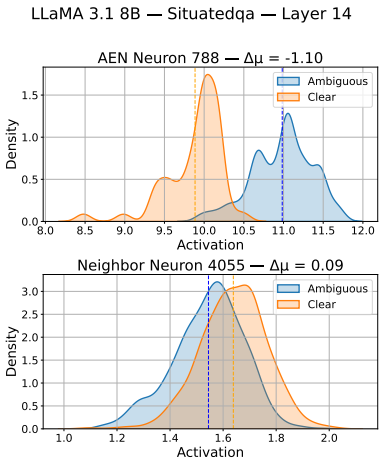
(a) Mistral 7B — AmbigQA



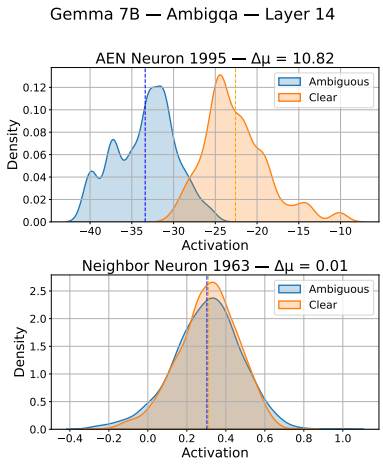
(b) Mistral 7B — SituatedQA



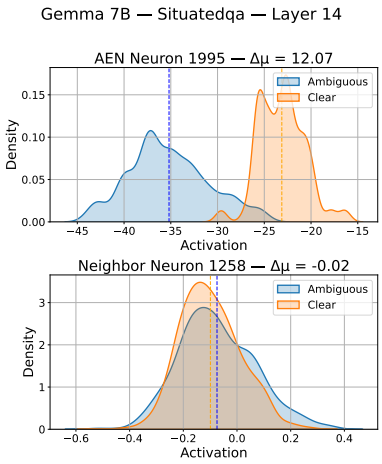
(c) LLaMA 3.1 8B — AmbigQA



(d) LLaMA 3.1 8B — SituatedQA

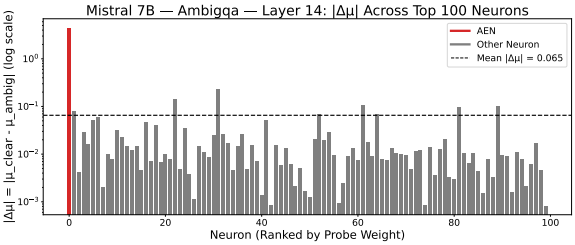


(e) Gemma 7B — AmbigQA

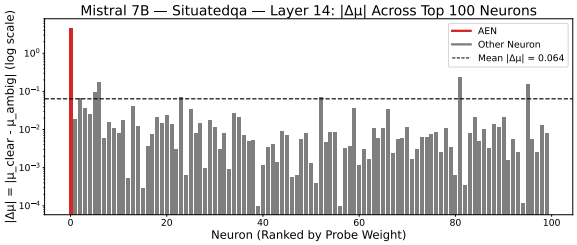


(f) Gemma 7B — SituatedQA

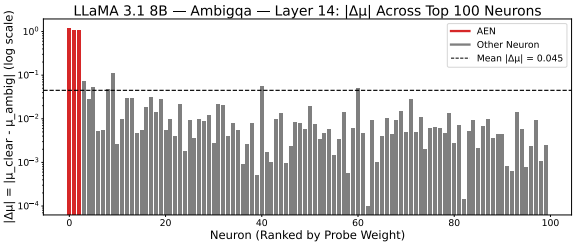
Figure 8: Activation distributions for AENs vs. neighbor neurons at Layer 14 across AMBIGQA and SITUATEDQA. Each row is a model; each column is a dataset. AENs show larger activation shifts between ambiguous and clear inputs.



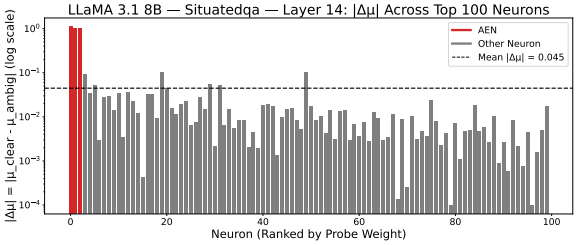
(a) Mistral 7B — AmbigQA



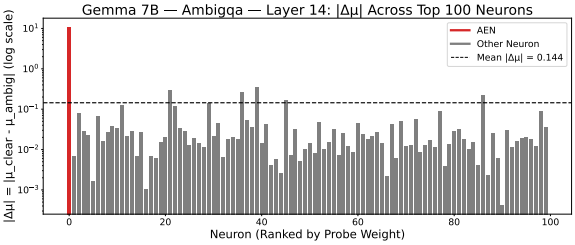
(b) Mistral 7B — SituatedQA



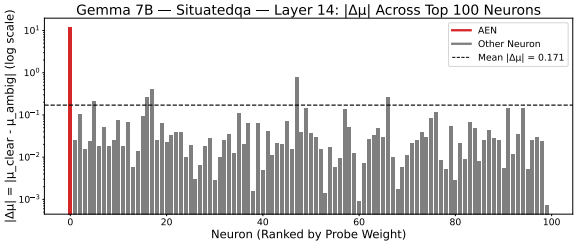
(c) LLaMA 3.1 8B — AmbigQA



(d) LLaMA 3.1 8B — SituatedQA



(e) Gemma 7B — AmbigQA



(f) Gemma 7B — SituatedQA

Figure 9: $|\Delta\mu|$ for the top-50 probe-weighted neurons on AMBIGQA and SITUATEDQA. In every model, AENs rank among the top positions and stand out from neighboring neurons.

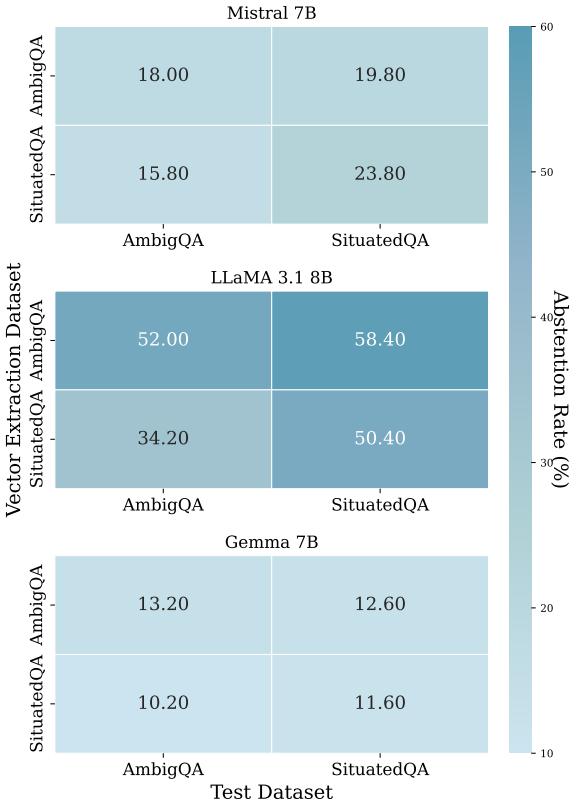


Figure 10: Cross-domain abstention rates with AEN-only steering. Rows correspond to the dataset used for extracting the ambiguity direction, and columns to the test set. AENs generalize across domains, especially in larger models like LLaMA 3.1 8B.