HALLUREFUSE: HALLUCINATION MITIGATION WITH KNOWLEDGE PRESERVING ACTIVATION STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) perform well in question answering, reasoning, and text generation. However, LLMs may produce plausible but factually incorrect answers when they lack sufficient knowledge or are uncertain. For reliable deployment, LLMs should answer known questions correctly while abstaining from unknown questions. Prior work has shown that activation steering can steer LLMs toward refusal without additional training, but existing methods typically apply interventions to all questions without distinguishing between known and unknown ones. Through empirical analysis, we show that indiscriminate steering induces a trade-off between reducing hallucinations on unknown questions and preserving performance on known questions. We further find that known question representations exhibit a compact spectral structure, revealing a low-dimensional subspace that can be protected. Based on these observations, we propose HALLUREFUSE, an activation intervention method that mitigates hallucinations while preserving known representations. HALLUREFUSE estimates the subspace of known representations and applies the intervention to the orthogonal subspace. The derived closed-form intervention guarantees that the protected known representations are unaffected. Extensive experiments across multiple LLMs and question answering benchmarks show that HALLUREFUSE mitigates hallucinations on unknown questions while achieving the highest average known accuracy across all LLMs.

1 INTRODUCTION

Large language models (LLMs) have shown strong abilities in question answering, reasoning, and text generation (Yehudai et al., 2026; Chen et al., 2026; Yuan et al., 2025). However, LLMs may still generate plausible but factually incorrect answers when the required knowledge is missing or uncertain, leading to hallucinations (Datta et al., 2026; Li et al., 2026). This problem is closely related to common training and evaluation objectives, which can reward confident guessing more than admitting uncertainty (Kalai et al., 2025; Huang et al., 2025). As a result, a reliable LLM should not only answer questions within its knowledge correctly, but also recognize questions it cannot answer reliably and abstain from answering them (Feng et al., 2024b; Zhang et al., 2024).

To enable LLMs to abstain when they cannot answer reliably, existing studies have developed training-based methods that explicitly optimize abstention behavior (Lee et al., 2026; Zhang et al., 2024; Li et al., 2025a). These methods use supervised fine-tuning, preference optimization, or reinforcement learning with refusal-aware data, knowledge boundary modeling, or uncertainty-aware objectives (Zhou et al., 2026; Zhai et al., 2026; Li et al., 2025b; Feng et al., 2024a). However, they require additional training data, optimization, and model updates, resulting in substantial computational costs. These limitations reduce the practicality of such methods for real-world deployment.

A recent study shows that internal activations of LLMs encode signals related to question answerability and can be steered through activation interventions to control refusal behavior without additional training (Lavi et al., 2026a). This finding suggests that activation steering can provide a training-free way to steer LLMs toward abstention when they cannot answer reliably. However, existing steering methods typically apply the intervention indiscriminately to all input queries, regardless of whether they are known or unknown to the LLM (You et al., 2026; Wang et al., 2025; Rimsky et al., 2024; Li et al., 2023). Such indiscriminate steering raises a key question of whether improving refusal on unknown questions may reduce performance on known questions.

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

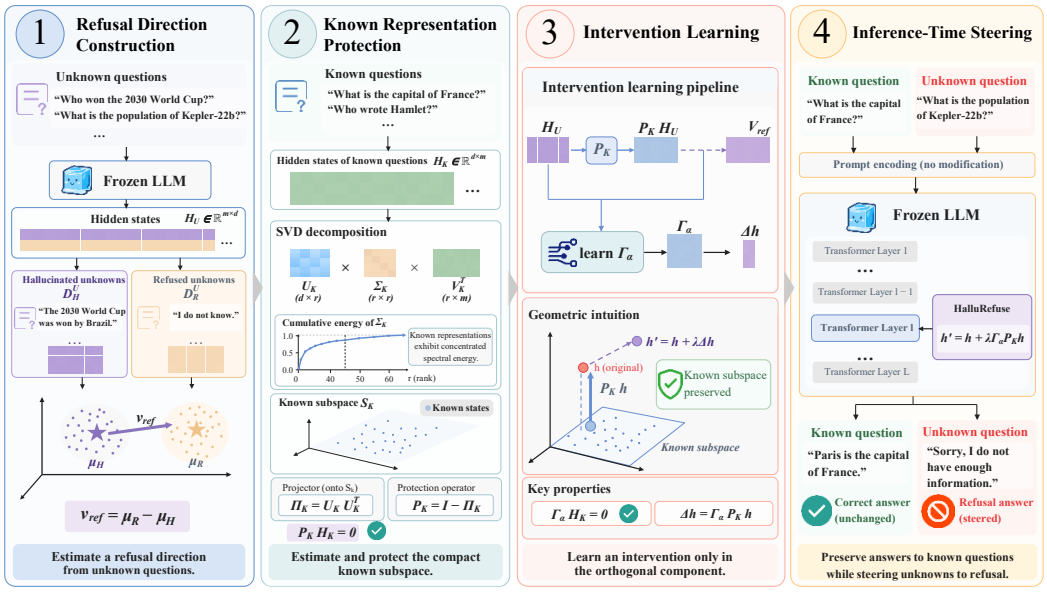


Figure 1: Overview of HALLUREFUSE. We construct a refusal direction, protect the compact subspace of known representations, and learn an intervention without affecting this protected subspace.

To examine the effects of indiscriminate steering, we conduct an empirical study summarized in Figure 2. Figure 2a shows a trade-off across four representative activation steering methods: stronger steering reduces hallucinations on unknown questions but also lowers accuracy on known questions. Figure 2b further shows that this accuracy degradation is associated with larger representation shifts on known questions, while Figure 2c visualizes such shifts under steering. Finally, Figure 2d reveals a compact spectral structure in known representations, with only 45 singular directions capturing **95.5%** of the total energy. Together, these observations motivate a central question: *Can we mitigate hallucinations on unknown questions while preserving the subspace of known representations?*

Motivated by these observations, we propose HALLUREFUSE, a training free activation intervention method that protects known representations during steering. We first construct a refusal direction from the representation difference between refused and hallucinated questions. We then estimate a compact subspace $\mathcal{S}_K$ from known representations and construct an orthogonal projector $\mathbf{P}_K$ onto $\mathcal{S}_K^{\perp}$. Next, we derive a regularized intervention matrix $\mathbf{\Gamma}_\alpha$ that maps the projected unknown representations toward the refusal direction. The resulting intervention admits a closed form solution and satisfies $\mathbf{\Gamma}_\alpha \mathbf{H}_K = \mathbf{0}$, ensuring zero intervention on the protected known representations. The theoretical guarantees are formally proved in Appendix D. At inference time, HALLUREFUSE applies the learned intervention $\mathbf{\Gamma}_\alpha \mathbf{P}_K$ to the hidden representations without updating the LLM parameters.

We evaluate HALLUREFUSE on multiple LLMs and question answering benchmarks. The results show that HALLUREFUSE achieves a better balance between hallucination mitigation on unknown questions and accuracy preservation on known questions than representative activation steering baselines. Across three LLMs and three QA benchmarks, HALLUREFUSE reduces the average hallucination rate on unknown questions by 5.88–14.48 percentage points over the original LLMs, while maintaining performance on known questions. Our main contributions are summarized as follows:

- We identify a consistent trade-off in activation steering for hallucination mitigation: stronger steering reduces hallucinations on unknown questions at the cost of accuracy on known questions.
- We show that this trade-off is associated with steering-induced shifts in known representations, and reveal that most spectral energy in known representations is concentrated in a compact subspace.
- We propose HALLUREFUSE, a training-free method that performs intervention in the subspace orthogonal to known representations. We derive a closed-form solution that theoretically guarantees zero intervention on known representations while steering unknown representations toward refusal.

2 PRELIMINARY

Dataset Format. Given a question answering dataset, each sample consists of a question $\mathbf{Q}$, its ground-truth answer $\mathbf{A}$, and a binary knowledge-state label y. The dataset is represented as

$$\mathcal{D} = \{(\mathbf{Q}_1, \mathbf{A}_1, y_1), \ldots, (\mathbf{Q}_n, \mathbf{A}_n, y_n)\}, \tag{1}$$

where n is the number of samples and $y_i \in \{0, 1\}$ indicates whether the target LLM knows the answer to $\mathbf{Q}_i$. Specifically, $y_i = 0$ denotes a known question, while $y_i = 1$ denotes an unknown question. The label is specific to the target LLM and does not represent an inherent property of the question. Following prior work (Yang et al., 2026), we sample multiple responses from the LLM for each question and determine its knowledge state based on the consistency of these responses.

Activation Steering. In this work, we focus on activation steering, an emerging and promising approach to mitigating hallucinations in LLMs (Yang et al., 2026; Lavi et al., 2026b). Given a question $\mathbf{Q}_i$ from $\mathcal{D}$, we extract the hidden state of last input token at layer l as the representation $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$, where d is the dimension. The core idea of activation steering is to modify this representation along a predefined refusal direction $\mathbf{r}^{(l)}$ during inference. Formally, the steering process is defined as

$$\mathbf{h}_i^{(l)'} \leftarrow \mathbf{h}_i^{(l)} + \lambda \mathbf{r}^{(l)}. \tag{2}$$

Here, the vector $\mathbf{r}^{(l)}$ captures the semantics of refusal behavior and guides the model toward refusing questions that it cannot answer reliably. The hyperparameter λ controls the steering strength.

Hallucination Mitigation. Following prior work (Yang et al., 2026; Zhang et al., 2024; Xu et al., 2024), we define a response as a refusal if the model explicitly states that it does not know the answer or cannot answer reliably. Given the generated response $\hat{\mathbf{A}}_i$ to question $\mathbf{Q}_i$, let $R(\hat{\mathbf{A}}_i) \in \{0, 1\}$ denote whether the response is a refusal, and $T(\hat{\mathbf{A}}_i, \mathbf{A}_i) \in \{0, 1\}$ denote whether it is truthful. For an unknown question, we define the hallucination indicator as

$$H_i = y_i\big(1 - R(\hat{\mathbf{A}}_i)\big)\big(1 - T(\hat{\mathbf{A}}_i, \mathbf{A}_i)\big), \tag{3}$$

where $H_i = 1$ indicates that the model provides an incorrect answer to an unknown question instead of refusing. Accordingly, the hallucination rate on unknown questions and the accuracy on known questions are defined as

$$\mathcal{H}_U = \frac{\sum_{i=1}^{n} y_i\big(1 - R(\hat{\mathbf{A}}_i)\big)\big(1 - T(\hat{\mathbf{A}}_i, \mathbf{A}_i)\big)}{\sum_{i=1}^{n} y_i}, \qquad \mathcal{A}_K = \frac{\sum_{i=1}^{n}(1 - y_i)T(\hat{\mathbf{A}}_i, \mathbf{A}_i)}{\sum_{i=1}^{n}(1 - y_i)}. \tag{4}$$

Due to space constraints, the related work is discussed in ***Appendix A***.

3 MOTIVATION

Recent methods based on activation steering for hallucination mitigation (You et al., 2026; Wang et al., 2025; Rimsky et al., 2024; Li et al., 2023) mainly improve the intervention by refining steering directions. However, they typically apply steering vectors indiscriminately to all input queries during inference, regardless of whether the query is known or unknown to the model. In fact, a reliable LLM should not only answer known questions correctly, but also recognize unknown questions and abstain from answering to reduce hallucinations (Feng et al., 2024b; Zhang et al., 2024). This motivates us to examine *how such indiscriminate steering affects known and unknown questions.*

3.1 OUTPUT-LEVEL EMPIRICAL OBSERVATIONS AND IMPLICATION

To examine the effects of indiscriminate steering, we vary the steering strength λ, which controls the magnitude of the intervention. Following the definitions in Section 2, we denote the hallucination rate on unknown questions and the accuracy on known questions under steering strength λ as $\mathcal{H}_U(\lambda)$ and $\mathcal{A}_K(\lambda)$, respectively. To quantify the changes relative to the unsteered LLM, we compute

$$\Delta\mathcal{H}_U(\lambda) = \mathcal{H}_U(0) - \mathcal{H}_U(\lambda), \qquad \Delta\mathcal{A}_K(\lambda) = \mathcal{A}_K(0) - \mathcal{A}_K(\lambda). \tag{5}$$

It is clear that larger $\Delta\mathcal{H}_U(\lambda)$ and $\Delta\mathcal{A}_K(\lambda)$ indicate greater hallucination mitigation and accuracy degradation, respectively. We conduct experiments on three LLMs by varying the strength λ across

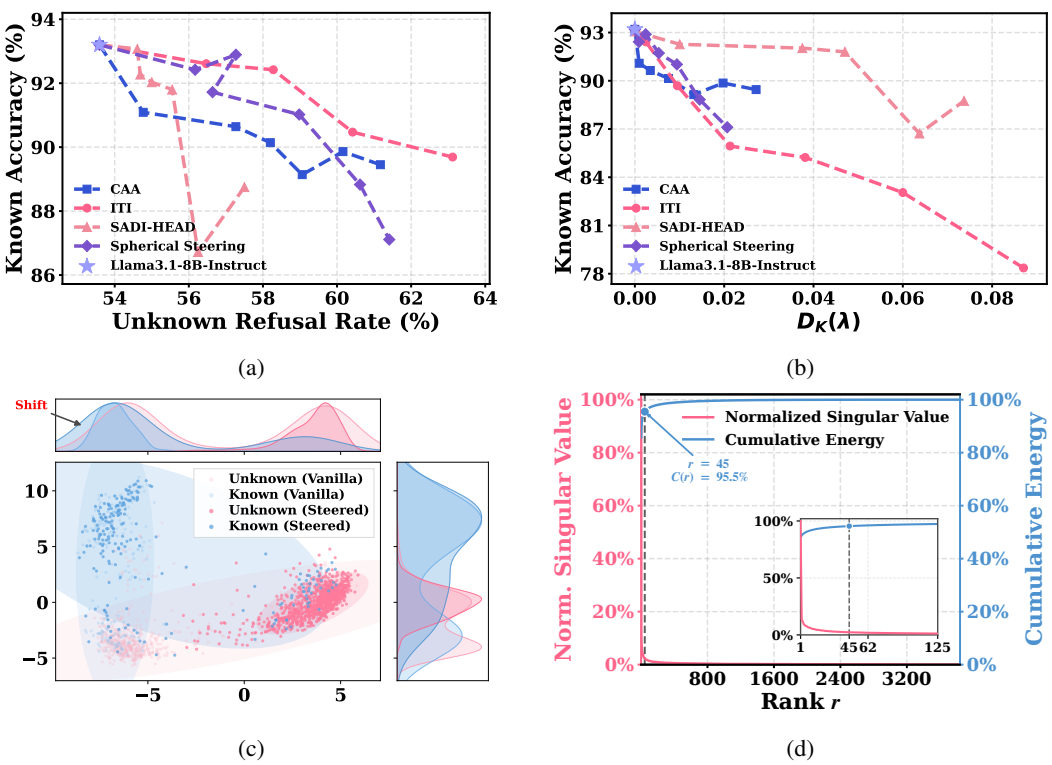


Figure 2: Motivation analysis of activation steering. (a) Trade-off between unknown refusal rate and known accuracy across steering methods. (b) Known accuracy versus steering-induced representation displacement $D_K(\lambda)$ across steering methods. (c) Representations of known and unknown questions before and after activation steering. (d) Singular-value spectrum and cumulative energy of known representations. More details of these empirical observations are provided in Appendix B.

multiple activation steering methods. For clarity, Figure 2a reports the results on LLaMA-3.1-8B-Instruct, while results on the other models and experimental details are provided in Appendix B.1.

Observation 1. As shown in Figure 2a, stronger steering generally leads to larger $\Delta\mathcal{H}_U(\lambda)$ and $\Delta\mathcal{A}_K(\lambda)$ across different methods. At the strongest evaluated setting of each method, $\Delta\mathcal{H}_U$ ranges from 3.91 to 9.53 percentage points, while $\Delta\mathcal{A}_K$ ranges from 3.51 to 6.09 percentage points. Taking CAA (Rimsky et al., 2024) as a representative activation steering method, increasing λ from 0 to 6 reduces $\mathcal{H}_U$ from 46.41% to 38.83%, while $\mathcal{A}_K$ decreases from 93.20% to 89.45%. We further observe similar trade-offs on three other representative methods, including ITI (Li et al., 2023), SADI-HEAD (Wang et al., 2025), and Spherical Steering (You et al., 2026). Despite minor fluctuations at individual steering strengths, the overall trend remains consistent across methods. The same trend is also observed on the other two LLMs, with detailed results provided in Appendix B.1.

Implication 1. These results show that stronger activation steering improves hallucination mitigation, but the improvement leads to a degradation in known-question performance. To understand the source of this trade-off, we next examine how steering changes the internal representations.

3.2 Representation-level Empirical Observations and Implication

To further analyze the trade-off, we examine representation shifts on known questions. Let $\mathcal{D}_K$ denote the set of known questions. We measure the representation shift induced by steering as

$$D_K(\lambda) = \frac{1}{|\mathcal{D}_K|} \sum_{i \in \mathcal{D}_K} \left(1 - \cos\left(\mathbf{h}_i^{(l)}(0), \mathbf{h}_i^{(l)}(\lambda)\right)\right), \tag{6}$$

where $\mathbf{h}_i^{(l)}(0)$ and $\mathbf{h}_i^{(l)}(\lambda)$ denote the last-token representations before and after steering at layer l, respectively. A larger $D_K(\lambda)$ indicates a larger representation shift on known questions. We conduct

this experiment on three LLMs with multiple activation steering methods. For clarity, Figure 2b reports the results on LLaMA-3.1-8B-Instruct, additional results are provided in Appendix B.2.

Observation 2. As shown in Figure 2b, a larger representation shift on known questions is generally associated with lower known accuracy across different steering methods. Specifically, for ITI (Li et al., 2023), as $D_K(\lambda)$ increases from 0 to approximately 0.09, $\mathcal{A}_K$ decreases from 93.20% to 78.36%, corresponding to a 14.84 percentage point drop. Similar trends are observed for three other representative methods, including CAA (Rimsky et al., 2024), SADI-HEAD (Wang et al., 2025), and Spherical Steering (You et al., 2026). Figure 2c provides a PCA visualization under CAA (Rimsky et al., 2024), showing a shift in known representations. More details are provided in Appendix B.2.

Implication 2. These results show that activation steering not only alters unknown representations, but also shifts the representations of known questions. Such representation shifts may explain the loss in known-question accuracy observed above. In Section 3.3, we further investigate whether known representations exhibit a compact structure that can be protected during steering.

3.3 Compact Subspace of Known Representations

The representation level results in Section 3.2 suggest that preserving known representations may help retain performance on known questions. We therefore examine whether these representations are concentrated in a compact subspace that can be protected during steering. Let $n_K = |\mathcal{D}_K|$, and let $\mathbf{H}_K^{(l)} \in \mathbb{R}^{n_K \times d}$ denote the hidden state matrix of known questions at layer l, where each row corresponds to one question representation. We apply SVD to the uncentered matrix:

$$\mathbf{H}_K^{(l)} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_q)$, $q = \min(n_K, d)$, and $\sigma_1 \geq \cdots \geq \sigma_q$. Inspired by (Wan et al., 2026), we measure the cumulative spectral energy captured by the top r singular directions as

$$C(r) = \sum_{i=1}^{r} \sigma_i^2 / \sum_{j=1}^{q} \sigma_j^2.$$

A larger $C(r)$ indicates that more spectral energy is captured by the top r directions. Figure 2d reports the results on LLaMA-3.1-8B-Instruct, with additional results provided in Appendix B.3.

Observation 3. As shown in Figure 2d, the spectral energy of known representations is highly concentrated. The top 45 singular directions capture 95.5% of the total energy, with $C(45) = 0.955$, indicating that most of the energy is concentrated in a small number of directions. A similar concentration pattern is observed on the other two LLMs, with detailed results provided in Appendix B.3.

Implication 3. The concentrated spectral structure suggests that known representations can be approximately characterized by a compact subspace. This provides a natural target for preservation during steering. These empirical observations motivate us to design a training-free method that mitigates hallucinations on unknown questions while preserving the known representation subspace.

4 Methodology

Motivated by the implications in Section 3.3, we develop HalluRefuse, a training-free method that performs intervention in the subspace orthogonal to known representations. This design mitigates hallucinations on unknown questions while preserving performance on known questions.

4.1 Refusal Direction Construction

We first construct a refusal direction to characterize the desired behavioral change for unknown questions. According to Eq. (3), we collect the unknown questions and further group them according to the model responses. Specifically, $\mathcal{D}_U^R$ contains unknown questions for which the model refuses to answer, whereas $\mathcal{D}_U^H$ contains unknown questions for which the model gives hallucinated responses. At layer l, we compute the mean representations of the two subsets as

$$\boldsymbol{\mu}_R^{(l)} = \frac{1}{|\mathcal{D}_U^R|} \sum_{i \in \mathcal{D}_U^R} \mathbf{h}_i^{(l)}, \qquad \boldsymbol{\mu}_H^{(l)} = \frac{1}{|\mathcal{D}_U^H|} \sum_{i \in \mathcal{D}_U^H} \mathbf{h}_i^{(l)},$$

where $\mathbf{h}_i^{(l)}$ is the representation of the i-th question at layer l. We then define the refusal direction as

$$\mathbf{v}_{\text{ref}}^{(l)} = \boldsymbol{\mu}_R^{(l)} - \boldsymbol{\mu}_H^{(l)}.$$

This direction captures the representation change from hallucination toward refusal and serves as the base steering direction. For simplicity, we omit the layer superscript in the following discussion.

4.2 Known Representation Protection

Observation 3 in Section 3.3 shows that known representations exhibit a compact low-dimensional structure. To avoid unnecessary changes to known representations during steering, we protect this subspace and restrict the intervention to its orthogonal complement. Specifically, we construct a projector that removes components aligned with the known representation subspace. Let $\mathbf{H}_K = [\mathbf{h}_1, \ldots, \mathbf{h}_n] \in \mathbb{R}^{d\times n}$ collect n known representations. We decompose the uncentered matrix as

$$\mathbf{H}_K = \mathbf{U}_K \boldsymbol{\Sigma}_K \mathbf{V}_K^\top, \tag{7}$$

where $\mathbf{U}_K \in \mathbb{R}^{d\times r_K}$, $\boldsymbol{\Sigma}_K \in \mathbb{R}^{r_K\times r_K}$, and $\mathbf{V}_K \in \mathbb{R}^{n\times r_K}$. The columns of $\mathbf{U}_K$ form an orthonormal basis of the protected subspace $\mathcal{S}_K = \text{col}(\mathbf{H}_K)$. We define the projector onto $\mathcal{S}_K$ and its complementary protection operator as

$$\mathbf{P}_K = \mathbf{I}_d - \mathbf{U}_K \mathbf{U}_K^\top. \tag{8}$$

Here, $\mathbf{I}_d$ is the d-dimensional identity matrix, and $\mathbf{P}_K \in \mathbb{R}^{d\times d}$ is the orthogonal projector onto $\mathcal{S}_K^\perp$. We next formalize this protection property in Lemma 1, which provides the basis for preserving known representations while constraining subsequent intervention to the orthogonal complement.

Lemma 1 (Exact protection of known representations)**.** *The matrix $\mathbf{P}_K$ is a rank-$(d-r_K)$ symmetric and idempotent projector. Moreover, $\mathbf{P}_K\mathbf{H}_K = \mathbf{0}$. Consequently, for every $\mathbf{s} \in \mathbb{R}^d$, $\mathbf{H}_K^\top \mathbf{P}_K \mathbf{s} = \mathbf{0}$. Equivalently, each known representation is orthogonal to any vector in the range of $\mathbf{P}_K$:*

$$\langle \mathbf{h}_i, \mathbf{P}_K \mathbf{s}\rangle = 0, \text{ for any } i = 1, \ldots, n.$$

Proof sketch. Since $\mathbf{U}_K^\top \mathbf{U}_K = \mathbf{I}_{r_K}$,

$$\mathbf{P}_K \mathbf{U}_K = \mathbf{U}_K - \mathbf{U}_K(\mathbf{U}_K^\top \mathbf{U}_K) = \mathbf{0}.$$

Substituting Eq. (7) gives

$$\mathbf{P}_K \mathbf{H}_K = (\mathbf{P}_K \mathbf{U}_K)\boldsymbol{\Sigma}_K \mathbf{V}_K^\top = \mathbf{0}.$$

Because $\mathbf{P}_K$ is symmetric, $\mathbf{H}_K^\top \mathbf{P}_K \mathbf{s} = (\mathbf{P}_K \mathbf{H}_K)^\top \mathbf{s} = \mathbf{0}$. Symmetry, idempotence, and the rank statement follow directly from Eq. (8). A complete derivation is provided in Appendix C. □

Remark. *Lemma 1 establishes that left multiplication by $\mathbf{P}_K$ nulls the known-representation matrix. Hence, the range of $\mathbf{P}_K$ contains no component along the protected subspace, providing geometric protection and restricting the intervention to directions orthogonal to known representations.*

4.3 Orthogonal Intervention

The protection matrix $\mathbf{P}_K$ from Section 4.2 satisfies $\mathbf{P}_K\mathbf{H}_K = \mathbf{0}$, which protects known questions from intervention. However, this protection does not encourage the model to refuse unknown questions. Therefore, we introduce an intervention matrix $\boldsymbol{\Gamma}_\alpha \in \mathbb{R}^{d\times d}$ to steer unknown questions toward refusal. Let $\mathbf{H}_U = [\mathbf{h}_1^U, \ldots, \mathbf{h}_{n_U}^U] \in \mathbb{R}^{d\times n_U}$ denote the representations of n_U unknown calibration questions. We use $\mathbf{v}_{\text{ref}}$ as the refusal direction and repeat it for all calibration questions:

$$\mathbf{V}_{\text{ref}} = \mathbf{v}_{\text{ref}} \mathbf{1}_{n_U}^\top \in \mathbb{R}^{d\times n_U}.$$

We seek an intervention matrix that maps the projected unknown representations $\mathbf{P}_K\mathbf{H}_U$ to the refusal target $\mathbf{V}_{\text{ref}}$. Since only $\boldsymbol{\Gamma}\mathbf{P}_K$ contributes to this mapping, we constrain $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\mathbf{P}_K$ so that $\boldsymbol{\Gamma}$ itself represents the effective intervention. We then solve the following optimization problem:

$$\boldsymbol{\Gamma}_\alpha = \underset{\boldsymbol{\Gamma}=\boldsymbol{\Gamma}\mathbf{P}_K}{\arg\min} \|\boldsymbol{\Gamma}\mathbf{P}_K\mathbf{H}_U - \mathbf{V}_{\text{ref}}\|_F^2 + \alpha \|\boldsymbol{\Gamma}\mathbf{P}_K\|_F^2. \tag{9}$$

where $\alpha > 0$ is a hyperparameter. The first term maps the protected unknown representations to the reference refusal direction. The second term regularizes the effective intervention using the Frobenius norm, which helps prevent overfitting. Importantly, this optimization problem admits a closed-form solution, as stated in Theorem 1. Implementation details are provided in Appendix F.

Table 1: Main results on three LLMs and three QA benchmarks. We report Known Accuracy, Known Refusal, and Unknown Hallucination (%). Higher known accuracy and lower refusal and hallucination rates are better. Best and second-best results are **bold** and underlined, respectively.

Method	Known Accuracy ↑			Known Refusal ↓			Unknown Hallucination ↓			Average		
	TriviaQA	EntityQA	PopQA	TriviaQA	EntityQA	PopQA	TriviaQA	EntityQA	PopQA	Known Accuracy	Known Refusal	Unknown Hallucination
LLaMA-3.1-8B-Instruct												
Baseline	**93.20**	86.88	86.09	9.61	12.97	18.12	46.41	21.64	23.02	88.72	13.57	30.36
5-shot ICL	90.16	70.00	80.77	6.72	19.53	13.02	51.95	16.48	28.00	80.31	13.09	32.14
ITI	91.88	86.26	86.97	8.12	**5.62**	**5.92**	42.27	20.47	22.92	88.37	6.55	28.55
CAA	91.09	86.41	86.09	10.23	6.72	17.16	45.23	19.45	20.86	87.86	11.37	28.51
SADI-HEAD	87.81	82.58	86.98	8.98	6.95	6.51	43.05	23.44	24.86	85.79	7.48	30.45
Spherical Steering	91.72	82.27	85.21	7.89	10.39	9.76	43.36	14.53	18.38	86.40	9.35	25.42
HalluRefuse (Ours)	92.27	**87.19**	**87.28**	**3.52**	6.02	8.28	**39.14**	**12.42**	**16.43**	**88.91**	**5.94**	**22.66**
Qwen2.5-7B-Instruct												
Baseline	94.45	86.91	88.89	1.48	3.52	**2.36**	72.19	59.30	59.45	90.08	**2.45**	63.65
5-shot ICL	**94.69**	84.19	85.86	1.56	4.33	5.05	69.45	48.44	53.69	88.25	3.65	57.19
ITI	91.25	84.79	86.53	1.41	4.33	7.07	69.30	47.58	47.52	87.52	4.27	54.80
CAA	93.44	86.20	87.88	1.88	3.52	4.04	71.48	56.17	54.60	89.17	3.15	60.75
SADI-HEAD	91.64	75.23	72.39	2.50	3.93	7.07	68.59	57.42	54.80	79.75	4.50	60.27
Spherical Steering	91.92	87.51	89.90	1.80	3.93	3.70	69.84	55.62	56.02	89.78	3.14	60.49
HalluRefuse (Ours)	94.06	**88.02**	**91.58**	**1.17**	**3.12**	3.37	**63.59**	**40.23**	**43.68**	**91.22**	2.55	**49.17**
Mistral-7B-Instruct-v0.3												
Baseline	96.41	91.48	92.46	**1.09**	**2.19**	3.41	69.22	41.72	55.82	93.45	2.23	55.59
5-shot ICL	94.77	86.88	84.91	1.17	4.77	6.33	65.86	39.06	**40.72**	88.85	4.09	**48.55**
ITI	95.16	90.63	92.21	1.64	3.91	**2.19**	68.75	38.45	56.06	92.67	2.58	54.42
CAA	95.70	90.94	90.75	1.17	2.73	3.61	66.95	41.41	54.09	92.46	2.50	54.15
SADI-HEAD	90.31	82.42	85.89	2.11	6.72	5.60	70.94	42.81	53.17	86.21	4.81	55.64
Spherical Steering	96.09	90.78	92.46	1.78	3.28	3.59	68.52	40.47	55.48	93.11	2.88	54.82
HalluRefuse (Ours)	**96.88**	**91.95**	**92.70**	1.17	2.89	**2.19**	**65.63**	**31.95**	51.56	**93.84**	**2.08**	49.71

Theorem 1 (Closed-form orthogonal intervention). *For any $\alpha > 0$, Eq. (9) has the unique solution*

$$\mathbf{\Gamma}_\alpha = \mathbf{V}_{\text{ref}} \mathbf{H}_U^\top \mathbf{P}_K \left(\mathbf{P}_K \mathbf{H}_U \mathbf{H}_U^\top \mathbf{P}_K + \alpha \mathbf{I}_d \right)^{-1}. \tag{10}$$

Moreover,

$$\mathbf{\Gamma}_\alpha \mathbf{P}_K = \mathbf{\Gamma}_\alpha, \qquad \mathbf{\Gamma}_\alpha \mathbf{H}_K = \mathbf{0}. \tag{11}$$

The first identity in Eq. (11) shows that $\mathbf{\Gamma}_\alpha$ acts only on representations projected by $\mathbf{P}_K$. The second ensures zero intervention on protected known representations. At inference time, we apply

$$\Delta \mathbf{h} = \mathbf{\Gamma}_\alpha \mathbf{P}_K \mathbf{h}, \qquad \mathbf{h}' = \mathbf{h} + \lambda \Delta \mathbf{h},$$

where λ controls the intervention strength. Eq. (9) encourages $\mathbf{\Gamma}_\alpha \mathbf{P}_K \mathbf{H}_U$ to map to $\mathbf{V}_{\text{ref}}$. When $\alpha > 0$, this mapping is regularized and approximate. The proof of Theorem 1 is provided in Appendix D. Figure 1 presents the overall framework and intervention pipeline of HALLUREFUSE.

5 EXPERIMENTS

In this section, we present empirical results to evaluate the effectiveness of HALLUREFUSE.

5.1 EXPERIMENTAL SETUP

Datasets and LLMs. We use three question answering benchmarks: TriviaQA (Joshi et al., 2017), EntityQA (Ferrando et al., 2025), and PopQA (Mallen et al., 2023). We evaluate three open-source LLMs: LLaMA-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3 (Grattafiori et al., 2024; Yang et al., 2024; Jiang et al., 2023). Following prior work (Yang et al., 2026), we partition each dataset into known and unknown questions. Additional details of the knowledge-state labeling, dataset sampling procedures, and LLM configurations are provided in Appendix E.

Baselines. We compare HALLUREFUSE with the original LLM, 5-shot in-context learning (ICL), and four activation steering methods. These methods include ITI (Li et al., 2023), CAA (Rimsky et al., 2024), SADI-HEAD (Wang et al., 2025), and Spherical Steering (You et al., 2026). Implementation details and hyperparameter settings for all baselines are provided in Appendix F.

Evaluation. We report known-question accuracy $\mathcal{A}_K$ and unknown-question hallucination rate $\mathcal{H}_U$, as defined in Eq. (4). A hallucination is an incorrect answer to an unknown question that the model

Table 2: Cross-dataset generalization results on three LLMs and three QA benchmarks. For each target dataset, the intervention is constructed using the other two datasets. The target dataset is not used during construction. Best and second-best results are **bold** and underlined, respectively.

Method	Known Accuracy ↑			Known Refusal ↓			Unknown Hallucination ↓			Average		
	TriviaQA	EntityQA	PopQA	TriviaQA	EntityQA	PopQA	TriviaQA	EntityQA	PopQA	Known Accuracy	Known Refusal	Unknown Hallucination
LLaMA-3.1-8B-Instruct												
Baseline	93.20	86.88	86.09	9.61	12.97	18.12	46.41	21.64	23.02	88.72	13.57	30.36
CAA	94.61	91.17	**92.31**	7.07	2.89	**2.07**	50.39	16.64	22.27	92.70	4.01	29.77
ITI	93.67	92.58	89.64	10.86	4.22	4.44	47.66	15.00	19.14	91.96	6.51	27.27
SADI-HEAD	**96.95**	88.91	90.24	5.00	**2.03**	8.28	51.02	18.91	19.78	92.03	5.10	29.90
Spherical Steering	92.42	92.27	92.01	6.25	4.61	6.21	44.61	14.92	18.70	92.23	5.69	26.08
HalluRefuse (Ours)	95.31	**93.28**	**92.31**	**4.38**	2.27	5.33	**40.63**	**13.98**	**16.32**	**93.63**	**3.99**	**23.64**
Qwen2.5-7B-Instruct												
Baseline	94.45	86.91	88.89	1.48	3.52	2.36	72.19	59.30	59.45	90.08	2.45	63.65
CAA	95.70	94.46	92.26	1.33	1.71	**1.35**	75.08	61.41	60.47	94.14	1.46	65.65
ITI	95.63	**94.66**	93.27	2.03	2.52	3.37	**67.73**	54.14	55.21	94.52	2.64	59.03
SADI-HEAD	82.50	84.99	86.87	4.06	3.93	1.68	73.44	52.03	78.36	84.79	3.22	67.94
Spherical Steering	95.94	92.85	94.28	**1.25**	**1.61**	**1.35**	73.20	60.70	61.17	94.36	**1.40**	65.02
HalluRefuse (Ours)	**97.27**	**94.66**	**94.95**	1.33	2.42	1.68	68.83	**48.36**	**52.68**	**95.63**	1.81	**56.62**
Mistral-7B-Instruct-v0.3												
Baseline	96.41	91.48	92.46	1.09	2.19	3.41	69.22	41.72	55.82	93.45	2.23	55.59
CAA	95.78	90.55	96.11	1.33	2.89	1.46	71.25	43.91	58.82	94.15	1.89	57.99
ITI	95.94	91.09	90.55	1.25	2.03	2.89	70.47	46.09	55.94	92.53	2.06	57.50
SADI-HEAD	96.56	87.73	89.54	2.11	3.28	2.92	70.47	47.50	58.82	91.28	2.77	58.93
Spherical Steering	95.70	89.84	96.11	1.88	3.67	1.46	67.81	44.14	56.06	93.88	2.34	56.00
HalluRefuse (Ours)	**97.97**	**96.80**	**96.35**	**0.94**	**1.02**	**1.22**	**62.58**	**40.23**	**51.79**	**97.04**	**1.06**	**51.53**

does not refuse. We also report Known Refusal to measure refusal behavior on known questions. Higher $\mathcal{A}_K$ and lower values of the other two metrics indicate better performance. All values are percentages. The Average columns give the arithmetic mean across the three datasets for each model.

5.2 Main Results

Table 1 shows that HalluRefuse achieves the lowest unknown hallucination rate in eight of the nine model and dataset combinations. It also achieves the highest average known accuracy on all three LLMs. Compared with the original models, it reduces average $\mathcal{H}_U$ by 7.70, 14.48, and 5.88 percentage points on LLaMA-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3 (Grattafiori et al., 2024; Yang et al., 2024; Jiang et al., 2023), respectively, while slightly improving average $\mathcal{A}_K$ by 0.19, 1.14, and 0.39 points. HalluRefuse also consistently outperforms all activation steering baselines on both average $\mathcal{H}_U$ and $\mathcal{A}_K$ across the three models. Although 5-shot ICL achieves a slightly lower average $\mathcal{H}_U$ on Mistral, it substantially reduces $\mathcal{A}_K$ to 88.85%, compared with 93.84% for HalluRefuse. Overall, these results demonstrate that HalluRefuse effectively reduces hallucinations while preserving performance on known questions.

5.3 Generalization Results

We evaluate cross-dataset generalization by constructing the intervention from two datasets and testing on the held-out third dataset, without using target data during construction. As shown in Table 2, HalluRefuse achieves the lowest $\mathcal{H}_U$ in eight of nine model-target pairs and the best average $\mathcal{A}_K$ and $\mathcal{H}_U$ on all three LLMs. Compared with the original LLMs, average $\mathcal{H}_U$ decreases by 6.72, 7.03, and 4.06 percentage points, while average $\mathcal{A}_K$ increases by 4.91, 5.55, and 3.59 points on LLaMA, Qwen, and Mistral, respectively. Further details are provided in Appendix G.

5.4 Ablation Studies

Effect of intervention strength. Figure 3(a) varies λ from 0 to 2. As λ increases, $\mathcal{H}_U$ gradually decreases from 46.41% to about 34%. Meanwhile, $\mathcal{A}_K$ stays close to 90% throughout the sweep. Known refusal also remains low for all positive values of λ. These results show that HalluRefuse can increase the intervention strength without causing a sharp drop in known questions performance. To balance hallucination mitigation and known questions performance, we set $\lambda = 1$ by default.

Effect of intervention subspace rank. Figure 3(b) varies r from 0 to 4096. A small intervention subspace reduces hallucinations while preserving known-question performance. At $r = 410$, $\mathcal{A}_K$

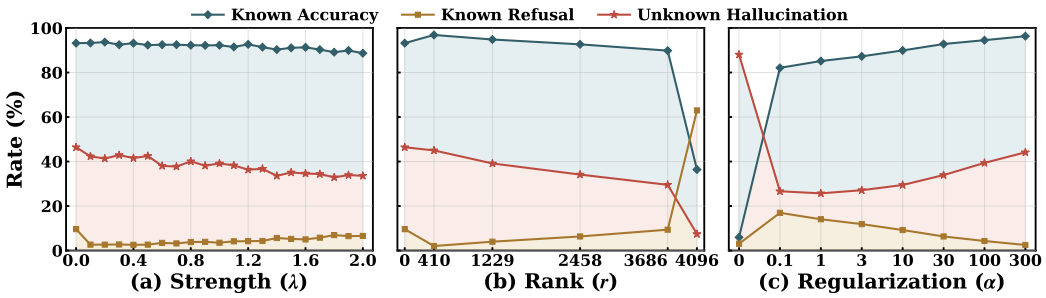


Figure 3: Ablation studies of HALLUREFUSE on LLaMA-3.1-8B-Instruct using the TriviaQA. We vary (a) intervention strength λ, (b) known subspace rank r, and (c) regularization coefficient α.

Unknown Question: In which decade of the 18th century did purpose built Prisoner of War Camps first appear?
ω/o **HalluRefuse:** Purpose-built prisoner of war camps … in the 1750s and 1760s.
ω **HalluRefuse:** I'm not sure in which ...

Known Question: Who became British Foreign Secretary in May 2010?
ω/o **HalluRefuse:** William Hague became the British Foreign Secretary in May 2010.
ω **HalluRefuse:** William Hague became the British Foreign Secretary in May 2010.

Figure 4: Case study of HALLUREFUSE. For the unknown question, the method replaces an incorrect answer with a refusal. For the known question, it keeps the correct answer unchanged.

improves from 93.20% to 96.88%, Known Refusal drops from 9.61% to 2.03%, and $\mathcal{H}_U$ slightly decreases. As r increases, $\mathcal{H}_U$ continues to decrease, but this comes with lower known accuracy and higher refusal on known questions. At full rank $r = 4096$, $\mathcal{H}_U$ drops to 7.42%, while $\mathcal{A}_K$ falls to 36.41% and Known Refusal rises to 62.97%, indicating over-refusal on known questions.

Effect of regularization. Figure 3(c) varies α from 0 to 300. Without regularization, $\mathcal{A}_K$ is only 5.94% and $\mathcal{H}_U$ reaches 88.05%. Setting $\alpha = 0.1$ improves $\mathcal{A}_K$ to 82.11% and reduces $\mathcal{H}_U$ to 26.64%. As α increases further, known accuracy continues to improve and Known Refusal decreases, while $\mathcal{H}_U$ reaches its minimum of 25.70% at $\alpha = 1$ and then gradually increases. To balance hallucination mitigation and known questions performance, we set $\alpha = 10$ by default.

5.5 CASE STUDY

Figure 4 compares model responses before and after applying HALLUREFUSE. For the unknown question, the original model gives a specific but incorrect answer instead of refusing. With HALLUREFUSE, the model states that it is unsure and avoids this hallucination. For the known question, both settings give the same correct answer, "William Hague". This contrast illustrates how HALLUREFUSE promotes refusal on unknown questions while preserving answers to known questions.

6 CONCLUSION

In this paper, we study hallucination mitigation through activation steering while preserving accuracy on known questions. We find that indiscriminate steering shifts known representations, which are concentrated in a compact subspace. We propose HALLUREFUSE, which protects the known representation subspace and steers its orthogonal complement toward refusal without training. Its closed form intervention leaves protected known representations unchanged. Results on three LLMs and three question answering benchmarks show lower hallucination rates on unknown questions while maintaining known accuracy. Ablations show how intervention strength, subspace rank, and regularization affect the balance between hallucination mitigation and known-question performance.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing and improving the clarity and readability of the manuscript. We have reviewed and revised all AI-assisted text and take responsibility for the final content of this work.

ETHICS STATEMENT

Our study adheres to the ICLR Code of Ethics. All experiments were conducted on publicly available datasets, as listed in Appendix E. The study does not involve human subjects, private or sensitive data, or personally identifiable information. We are not aware of any additional ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

All models and benchmark datasets employed in this study are publicly available. Detailed descriptions of the datasets are given in **Appendix E**, while the implementation details of our method are provided in **Appendix F**. To ensure reproducibility, all experiments were conducted on two NVIDIA L40 GPUs within a controlled environment, using Python 3.11.15 and PyTorch 2.6.0.

REFERENCES

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

Jiangjie Chen, Qianyu He, Siyu Yuan, Aili Chen, Zhicheng Cai, Weinan Dai, Hongli Yu, Jiaze Chen, Xuefeng Li, Qiying Yu, et al. Enigmata: Scaling logical reasoning in large language models with synthetic verifiable puzzles. *NeurIPS*, 2026.

Debtanu Datta, Mohan Kishore Chilukuri, Yash Kumar, Saptarshi Ghosh, and Muhammad Bilal Zafar. Do llm hallucination detectors suffer from low-resource effect? In *ACL*, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. 2024.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-llm collaboration. In *ACL*, 2024a.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-llm collaboration. In *ACL (1)*, pp. 14664–14690. Association for Computational Linguistics, 2024b.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *ICLR*, 2025.

Tim Franzmeyer, Archie Sravankumar, Lijuan Liu, Yuning Mao, Rui Hou, Sinong Wang, Jakob Foerster, Luke Zettlemoyer, and Madian Khabsa. High accuracy, less talk (halt): Reliable llms through capability-aligned finetuning. In *ICLR*, 2026.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *ICLR*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, and Bing Qin. Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning. In *ACL*, 2025.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In *EMNLP*, 2025.

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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *CoRR*, abs/2509.04664, 2025. doi: 10.48550/arXiv.2509.04664. URL https://doi.org/10.48550/arXiv.2509.04664.

Maor Juliet Lavi, Tova Milo, and Mor Geva. Detecting (un) answerability in large language models with linear directions. In *EACL*, 2026a.

Maor Juliet Lavi, Tova Milo, and Mor Geva. Detecting (un)answerability in large language models with linear directions. In *EACL*, 2026b.

Joosung Lee, Hwiyeol Jo, Donghyeon Ko, Kyubyung Chae, Cheonbok Park, and Jeonghoon Kim. What models know, how well they know it: Knowledge-weighted fine-tuning for learning when to say "i don't know". *CoRR*, abs/2604.05779, 2026.

Jiaqi Li, Yixuan Tang, and Yi Yang. Know the unknown: An uncertainty-sensitive method for LLM instruction tuning. In *ACL*, 2025a.

Jiaqi Li, Yixuan Tang, and Yi Yang. Know the unknown: An uncertainty-sensitive method for llm instruction tuning. In *ACL*, 2025b.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *NeurIPS*, 2023.

Xinyi Li, Zhen Fang, Yongxin Deng, Jinyuan Luo, Hongnan Ma, Changdae Oh, Zijing Shi, Shanshan Ye, Hanchen Wang, Shu-Lin Chen, et al. Openhaldet: A unified benchmark for hallucination detection across diverse generation scenarios. *arXiv preprint arXiv:2606.06959*, 2026.

Jinyuan Luo, Zhen Fang, Yixuan Li, Seongheon Park, and Ling Chen. Shaking to reveal: Perturbation-based detection of llm hallucinations. *arXiv preprint arXiv:2506.02696*, 2025.

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

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. 2024.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *NeurIPS*, 2024.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. Spectral editing of activations for large language model alignment. 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *ACL*, 2024.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. *NeurIPS*, 2024.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248, 2023.

Zhang Wan, Yu Li, Tianze Huang, Haochen Li, Juan Cao, and Sheng Tang. Video-svd: Efficient video diffusion via orthogonal basis composition. In *ICML*, 2026.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors. In *ICLR*, 2025.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Wannan Yang, Xinchi Qiu, Lei Yu, Yuchen Zhang, Aobo Yang, Narine Kokhlikyan, Nicola Cancedda, and Diego Garcia-Olano. Hallucination reduction with casal: Contrastive activation steering for amortized learning. In *ICLR*, volume 2026, pp. 10611–10665, 2026.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. A survey on evaluation of llm-based agents. In *ACL*, 2026.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical steering: Geometry-aware activation rotation for language models. 2026.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Boji Shan, Zeyuan Liu, Jia Deng, Huimin Chen, Ruobing Xie, et al. Advancing llm reasoning generalists with preference trees. In *ICLR*, 2025.

Haotian Zhai, Jingcheng Liang, and Dongyeop Kang. Abstain-r1: Calibrated abstention and post-refusal clarification via verifiable rl. In *ACL*, 2026.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'i don't know'. In *NAACL*, 2024.

Xinyu Zhou, Chang Jin, Carsten Eickhoff, Zhijiang Guo, and Seyed Ali Bahrainian. When silence is golden: Can llms learn to abstain in temporal QA and beyond? *ICLR*, 2026.

A Related Work

Large Language Models. Large language models (LLMs) have achieved strong performance in text understanding, question answering, reasoning, and generation (Chen et al., 2026; Yuan et al., 2025; Du et al., 2024). Despite these advances, LLMs may still generate plausible but factually incorrect responses when relevant knowledge is absent or insufficient, leading to hallucinations (Luo et al., 2025; Sriramanan et al., 2024). Prior work has studied this problem from different perspectives, including factuality verification, uncertainty estimation, retrieval augmentation, and behavioral control. A central challenge is that models are often encouraged to provide an answer even when their internal knowledge is insufficient. As a result, reducing hallucinations requires not only improving factual generation, but also enabling models to recognize when they should refrain from answering.

Learning to Abstain. Abstention provides a direct way to reduce hallucinations by allowing a model to refuse questions for which it lacks sufficient knowledge. Existing studies improve abstention through uncertainty estimation, confidence calibration, refusal learning, and explicit supervision on known and unknown examples (Mohri & Hashimoto, 2024; Franzmeyer et al., 2026; Nikitin et al., 2024). These methods aim to distinguish answerable inputs from those that should trigger refusal, thereby reducing incorrect responses on uncertain or unknown questions. However, effective abstention requires a careful balance: a model should refuse when its knowledge is insufficient while preserving its ability to answer questions it already knows. Over refusal can therefore be as undesirable as under-refusal, since excessive abstention reduces the utility of the model on otherwise answerable inputs. In this work, we focus on this balance and study how to improve refusal on unknown questions without unnecessarily degrading performance on known questions.

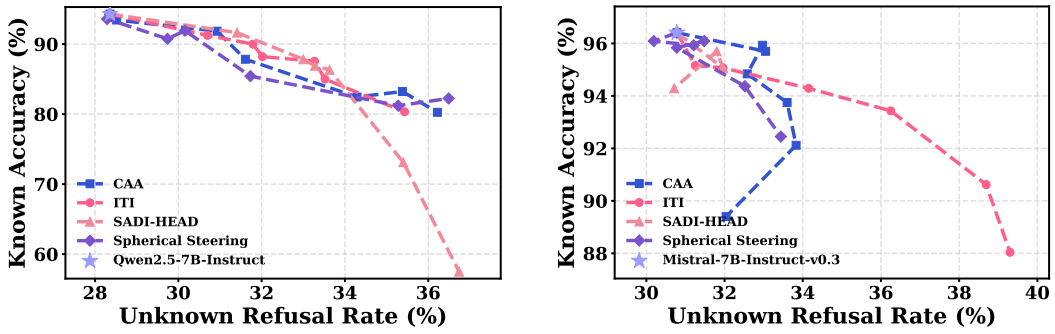
Activation Steering. Activation steering provides a lightweight inference-time approach for controlling model behavior by modifying hidden representations during the forward pass without updating model parameters. Existing methods typically identify behavior-related directions and apply them to intermediate activations to promote refusal, factuality, or other desired behaviors (Ji et al., 2025; Rimsky et al., 2024; Turner et al., 2023). Such steering directions can be constructed from supervised probes, differences between mean activations of contrastive samples, behavior-related representation subspaces, or features extracted by sparse autoencoders (Ferrando et al., 2025; Arditi et al., 2024; Li et al., 2023; Qiu et al., 2024). Compared with additional fine-tuning, activation steering is flexible and can be applied directly at inference time. However, most existing approaches apply a similar intervention across inputs without explicitly accounting for whether a question is already known to the model. Consequently, directly adding steering vectors may perturb representations that do not require intervention, shifting them away from their original distribution and degrading performance on known inputs (Grant et al., 2026). Our work addresses this limitation by explicitly preserving the representation subspace associated with known questions while steering unknown questions toward refusal.

B Additional Details of Empirical Observations

B.1 Output-level Empirical Observations

We provide additional implementation details and results for the output-level analysis in Section 3. All experiments are conducted on TriviaQA using three LLMs, including LLaMA-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3. We evaluate four representative activation steering methods: ITI (Li et al., 2023), CAA (Rimsky et al., 2024), SADI-HEAD (Wang et al., 2025), and Spherical Steering (You et al., 2026). For ITI, CAA, and SADI-HEAD, we vary the steering strength λ from 1 to 6 with an interval of 1. For the LLaMA-3.1-8B-Instruct experiment, we use 48 intervention heads for ITI, apply CAA at layer 20, and use 4 intervention heads for SADI-HEAD based on a preliminary hyperparameter search. For Spherical Steering, the intervention strength is controlled by α. We vary α over $\{0.025, 0.05, 0.075, 0.1, 0.125, 0.15\}$ while fixing the intervention layer to 22, $\kappa = 20$, and $\beta = 0.1$.

We evaluate the full strength range for each method. However, excessively strong steering does not always produce stronger refusal behavior. For some methods, further increasing the intervention strength can disturb the original generation process and lead to additional incorrect responses. Therefore, Figure 2a reports the effective strength range before such degradation becomes appar-



(a) Qwen2.5-7B-Instruct (b) Mistral-7B-Instruct-v0.3

Figure 5: Trade-off between unknown refusal rate and known-question accuracy on two additional LLMs. Each curve corresponds to one activation steering method under different intervention strengths. Despite fluctuations at individual strengths, stronger refusal behavior is generally associated with lower known-question accuracy.

ent. This allows us to focus on the trade-off between hallucination mitigation and known-question performance rather than general generation degradation caused by overly strong intervention.

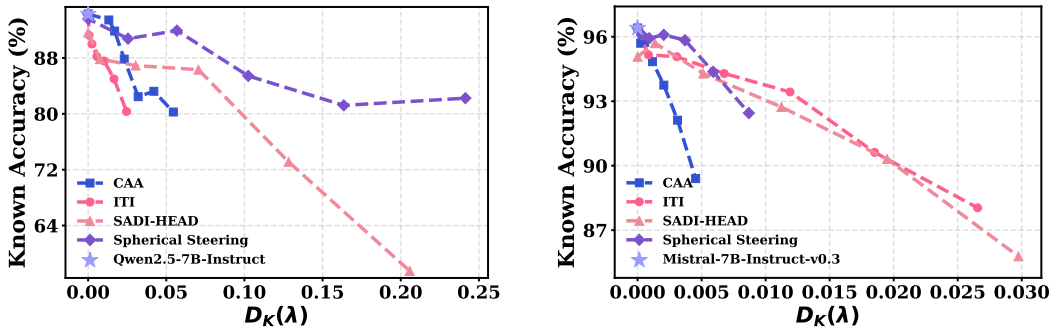
Figure 5 reports the results on the two additional LLMs. For visualization, we plot the unknown refusal rate against known-question accuracy under different steering strengths. On Qwen2.5-7B-Instruct, stronger steering generally increases the unknown refusal rate while reducing known-question accuracy across all four methods. The degradation is particularly pronounced for SADI-HEAD at stronger intervention strengths. A similar overall pattern is observed on Mistral-7B-Instruct-v0.3. Although several methods exhibit fluctuations at individual strengths, higher unknown refusal rates are generally associated with lower known-question accuracy. These additional results are consistent with Observation 1 in the main text. The trade-off between hallucination mitigation and known-question performance is therefore observed across different LLMs and multiple activation steering methods, rather than being specific to LLaMA-3.1-8B-Instruct.

B.2 Representation-level Empirical Observations

We extend the representation analysis in Figure 2b and Figure 2c along three axes. We study three LLMs on TriviaQA, four activation steering methods, and three question answering datasets on LLaMA-3.1-8B-Instruct. The evaluated methods are CAA (Rimsky et al., 2024), ITI (Li et al., 2023), SADI-HEAD (Wang et al., 2025), and Spherical Steering (You et al., 2026). This analysis tests whether the representation shift observed in the main text persists across models, methods, and datasets. For each question, we generate one response without steering and one response with steering under the same generation setting. We then concatenate the chat-formatted question and response and run a second forward pass. We extract the hidden state of the final non-padding token at the evaluation layer. This gives paired representations before and after steering. We compute $D_K(\lambda)$ on known questions using the definition in the main text.

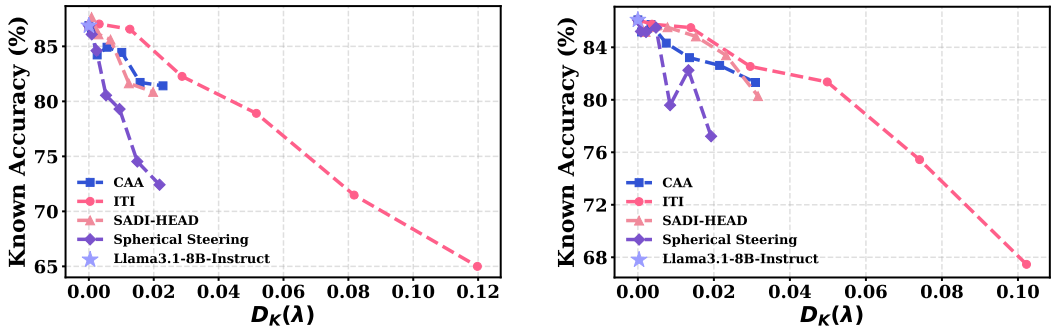
Figure 6 first extends the $D_K(\lambda)$ analysis from LLaMA-3.1-8B-Instruct to Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3 on TriviaQA. Together with Figure 2b, the results show the same relation on all three LLMs. Known accuracy generally decreases as the representation displacement grows. This trend is clear on Mistral-7B-Instruct-v0.3. Qwen2.5-7B-Instruct shows several local fluctuations, but its overall trend is unchanged.

We next test the same relation across datasets. Figure 7 reports results on EntityQA and PopQA using LLaMA-3.1-8B-Instruct. The four methods begin near the unsteered model and move toward lower known accuracy as $D_K(\lambda)$ increases. ITI spans the largest displacement range on both datasets and also produces the largest accuracy loss. CAA and SADI-HEAD cover smaller ranges, but they follow the same overall pattern. Spherical Steering has local fluctuations on PopQA. Its stronger settings still produce larger displacement and lower known accuracy. The link between representation displacement and accuracy loss therefore holds on all three evaluated datasets.



(a) Qwen2.5-7B-Instruct (b) Mistral-7B-Instruct-v0.3

Figure 6: Known accuracy versus the steering-induced displacement $D_K(\lambda)$ on two additional LLMs using TriviaQA. Each curve varies the intervention strength of one steering method. Larger displacement is generally associated with lower known accuracy.



(a) EntityQA (b) PopQA

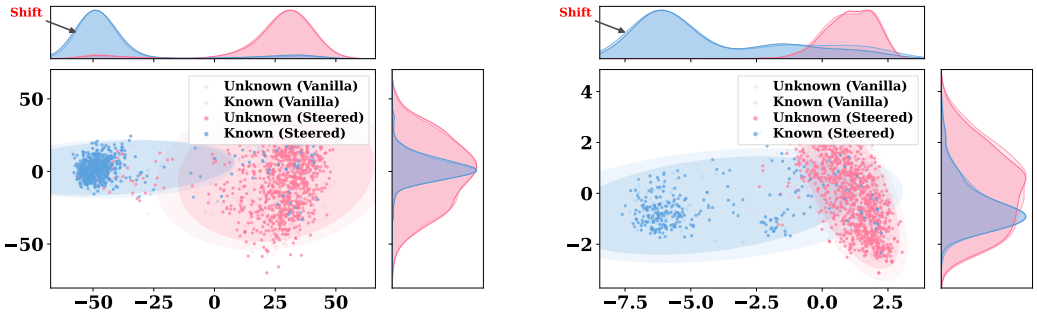
Figure 7: Known accuracy versus the steering-induced displacement $D_K(\lambda)$ on two additional datasets using LLaMA-3.1-8B-Instruct. Larger shifts generally coincide with larger losses in known accuracy across all four steering methods.

We also visualize the representation changes with PCA. Each panel contains the known and unknown representations before and after steering. Figure 8 compares CAA across LLMs on TriviaQA. The LLaMA-3.1-8B-Instruct result appears in Figure 2c. The Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3 results show the same behavior. CAA changes both the position and the shape of the known representation distribution. Thus, its effect is not confined to the unknown questions that the intervention aims to steer.

Figure 9 then compares different methods on LLaMA-3.1-8B-Instruct using TriviaQA. ITI, SADI-HEAD, and Spherical Steering all move the known representations away from their original distribution. Together with the CAA result in Figure 2c, this shows that the effect is shared by methods with different intervention rules.

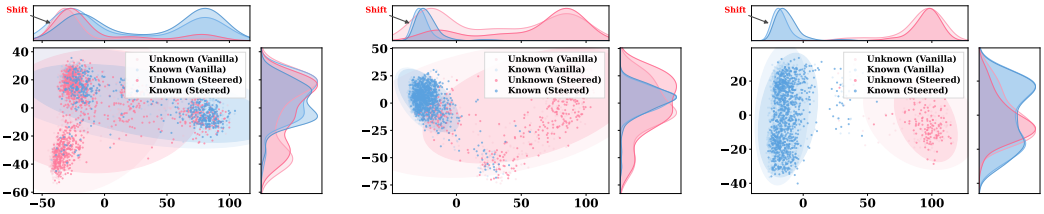
The new PCA results extend this method comparison to EntityQA and PopQA. Figure 10 shows clear changes in the known distribution under all four methods on EntityQA. The shift can change both the location and spread of the distribution. Figure 11 shows the same pattern on PopQA. The detailed geometry differs across methods, but the known representations are affected in every panel. These PCA views provide qualitative evidence. The quantitative size of each change is measured by $D_K(\lambda)$ in Figure 7.

Taken together, the quantitative and qualitative results show a stable pattern. Activation steering changes known representations across model families, intervention rules, and datasets. Larger changes are generally linked to lower known accuracy. This evidence motivates protecting the subspace of known representations while steering the remaining component.



(a) Qwen2.5-7B-Instruct (b) Mistral-7B-Instruct-v0.3

Figure 8: PCA views of known and unknown representations before and after CAA on two additional LLMs using TriviaQA. The result for LLaMA-3.1-8B-Instruct is shown in Figure 2c. CAA visibly shifts known representations on all three LLMs.



(a) ITI (b) SADI-HEAD (c) Spherical Steering

Figure 9: PCA views of known and unknown representations before and after three activation steering methods on LLaMA-3.1-8B-Instruct using TriviaQA. The CAA result is shown in Figure 2c. Each method visibly changes the known representation distribution.

B.3 Compact Subspace of Known Representations

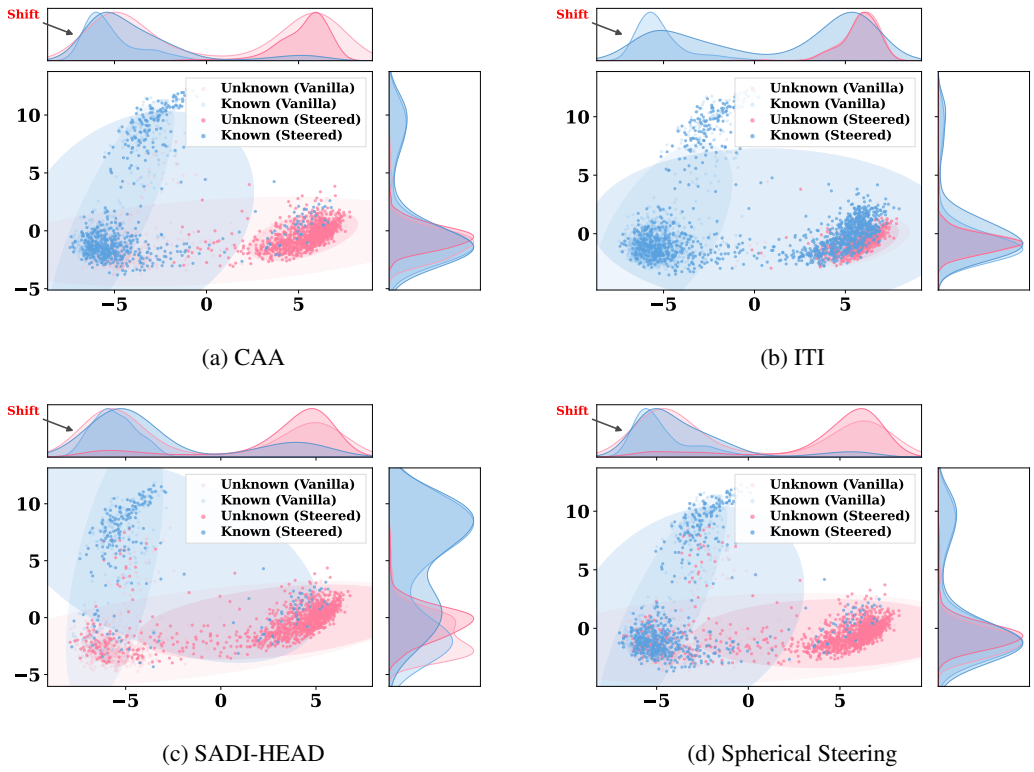
We extend the analysis in Section 3.3 to three question answering datasets and three LLMs. For each model and dataset, we collect the known questions and apply the model chat template with a generation prompt. We extract the hidden state of the final prompt token at intervention layer $l = 19$. Stacking these states as rows gives $\mathbf{H}_K^{(l)} \in \mathbb{R}^{n_K \times d}$, where n_K is the number of known questions and d is the hidden dimension. We compute the singular values of this matrix without centering, using float32 precision.

Let $\sigma_1 \geq \cdots \geq \sigma_q$ be the singular values, where $q = \min(n_K, d)$. The normalized energy of direction i and the cumulative energy of the first r directions are

$$p_i = \frac{\sigma_i^2}{\sum_{j=1}^{q} \sigma_j^2}, \qquad C(r) = \sum_{i=1}^{r} p_i.$$

Thus, $C(r)$ measures the fraction of the matrix energy captured by the first r singular directions. We use $r = 45$ as a common reference point across all model and dataset pairs. This choice follows the TriviaQA result for LLaMA-3.1-8B-Instruct in Figure 2d.

Table 3 summarizes the nine spectra. Across all pairs, the first 45 directions capture between 92.8% and 99.0% of the energy. EntityQA exhibits the highest concentration for all three models, while the lowest-concentration dataset varies across models. Figure 2d shows the LLaMA-3.1-8B-Instruct result on TriviaQA. Figure 12 shows the other two models on TriviaQA. Figures 13, 14, and 15 show the results on EntityQA and PopQA. The spectra vary across datasets and models, but the leading directions contain most of the uncentered energy in every setting.



(a) CAA (b) ITI

(c) SADI-HEAD (d) Spherical Steering

Figure 10: PCA views of known and unknown representations before and after four activation steering methods on EntityQA using LLaMA-3.1-8B-Instruct. Every method visibly changes the distribution of known representations.

Table 3: Cumulative energy $C(45)$ of known representations. Each value is computed from the corresponding model and dataset.

Dataset	LLaMA-3.1-8B-Instruction	Qwen2.5-7B-Instruct	Mistral-7B-Instruct-v0.3
TriviaQA	95.5%	96.1%	95.2%
EntityQA	97.9%	99.0%	96.7%
PopQA	95.8%	98.1%	92.8%

C Proof of Known Representation Protection

Proof of Lemma 1. The columns of $\mathbf{U}_K$ are orthonormal, so $\mathbf{U}_K^\top \mathbf{U}_K = \mathbf{I}_{r_K}$. The transpose of the protection operator is

$$\mathbf{P}_K^\top = (\mathbf{I}_d - \mathbf{U}_K \mathbf{U}_K^\top)^\top = \mathbf{I}_d - \mathbf{U}_K \mathbf{U}_K^\top = \mathbf{P}_K. \tag{12}$$

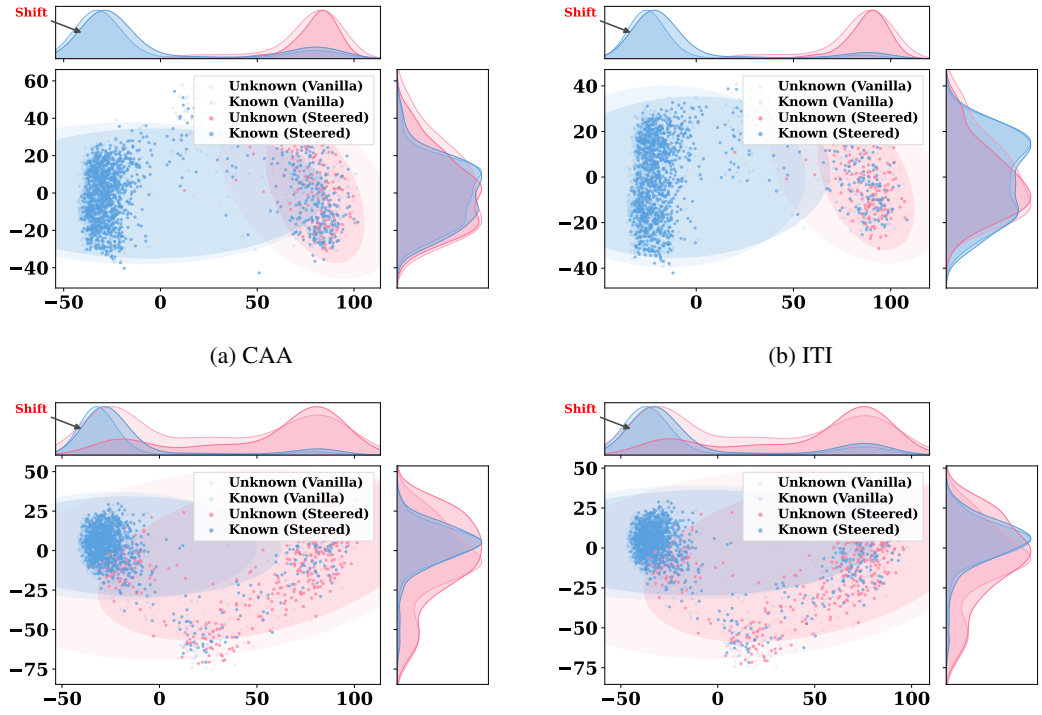
Its square is

$$\mathbf{P}_K^2 = (\mathbf{I}_d - \mathbf{U}_K \mathbf{U}_K^\top)^2 \tag{13}$$

$$= \mathbf{I}_d - 2\mathbf{U}_K \mathbf{U}_K^\top + \mathbf{U}_K (\mathbf{U}_K^\top \mathbf{U}_K) \mathbf{U}_K^\top \tag{14}$$

$$= \mathbf{I}_d - \mathbf{U}_K \mathbf{U}_K^\top = \mathbf{P}_K. \tag{15}$$

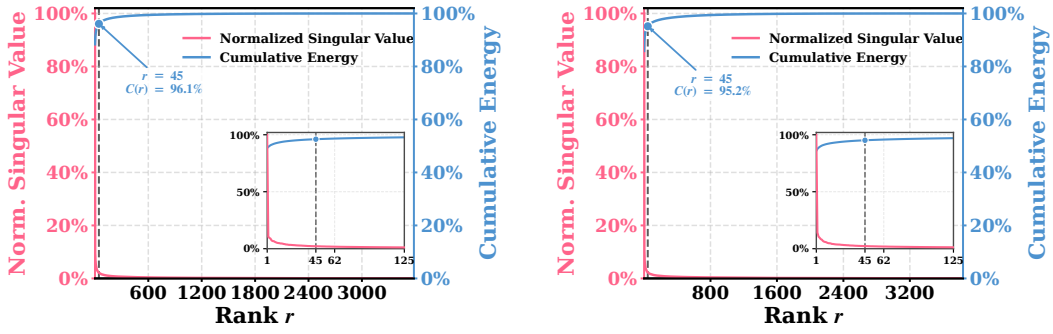
Thus, $\mathbf{P}_K$ is a symmetric idempotent matrix and therefore the orthogonal projector onto $\mathrm{span}(\mathbf{U}_K)^\perp$. Since $\dim \mathrm{span}(\mathbf{U}_K) = r_K$, we also have $\mathrm{rank}(\mathbf{P}_K) = d - r_K$.



(a) CAA (b) ITI

(c) SADI-HEAD (d) Spherical Steering

Figure 11: PCA views of known and unknown representations before and after four activation steering methods on PopQA using LLaMA-3.1-8B-Instruct. The known representation distribution changes under every method.



(a) Qwen2.5-7B-Instruct (b) Mistral-7B-Instruct-v0.3

Figure 12: Singular value spectra and cumulative energy of known representations on TriviaQA. The marker in each panel indicates $C(45)$.

We next establish the left-null protection identity. Direct expansion gives

$$\mathbf{P}_K\mathbf{U}_K = (\mathbf{I}_d - \mathbf{U}_K\mathbf{U}_K^\top)\mathbf{U}_K \tag{16}$$

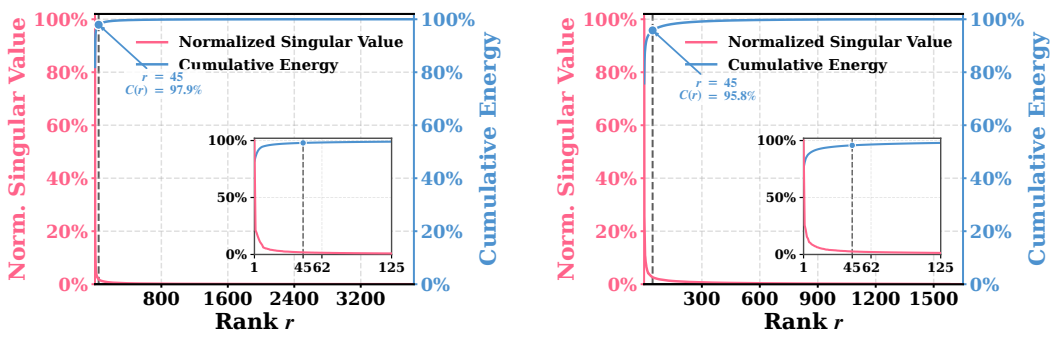
$$= \mathbf{U}_K - \mathbf{U}_K(\mathbf{U}_K^\top\mathbf{U}_K) \tag{17}$$

$$= \mathbf{0}. \tag{18}$$

Using the compact SVD of $\mathbf{H}_K$, it follows that

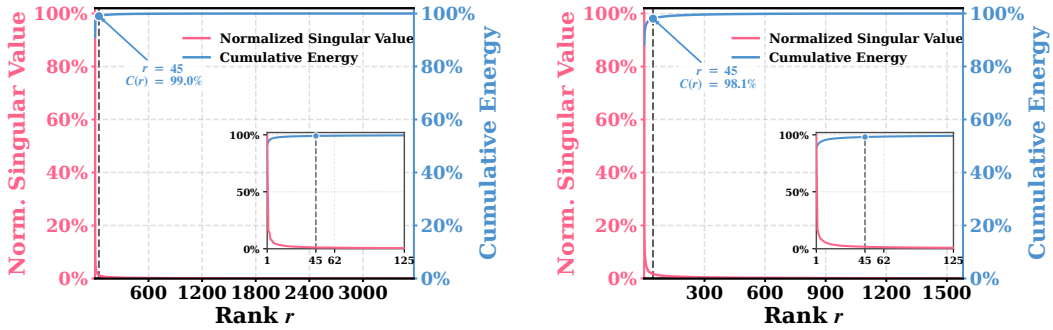
$$\mathbf{P}_K\mathbf{H}_K = \mathbf{P}_K\mathbf{U}_K\mathbf{\Sigma}_K\mathbf{V}_K^\top \tag{19}$$

$$= \mathbf{0}. \tag{20}$$



(a) EntityQA (b) PopQA

Figure 13: Singular value spectra and cumulative energy of known representations on LLaMA-3.1-8B-Instruct. The marker in each panel indicates $C(45)$.



(a) EntityQA (b) PopQA

Figure 14: Singular value spectra and cumulative energy of known representations on Qwen2.5-7B-Instruct. The marker in each panel indicates $C(45)$.

This identity also holds column-wise. Since $\mathbf{H}_K = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$,

$$\mathbf{P}_K \mathbf{h}_i = \mathbf{0}, \qquad i = 1, \ldots, n. \tag{21}$$

Finally, let $\mathbf{s} \in \mathbb{R}^d$ be arbitrary. The symmetry of $\mathbf{P}_K$ and Eq. (20) yield

$$\mathbf{H}_K^\top \mathbf{P}_K \mathbf{s} = \mathbf{H}_K^\top \mathbf{P}_K^\top \mathbf{s} \tag{22}$$

$$= (\mathbf{P}_K \mathbf{H}_K)^\top \mathbf{s} \tag{23}$$

$$= \mathbf{0}. \tag{24}$$

The i-th entry of this vector equality is

$$\mathbf{h}_i^\top \mathbf{P}_K \mathbf{s} = \langle \mathbf{h}_i, \mathbf{P}_K \mathbf{s} \rangle = 0, \tag{25}$$

which proves zero linear overlap with every protected known representation. □

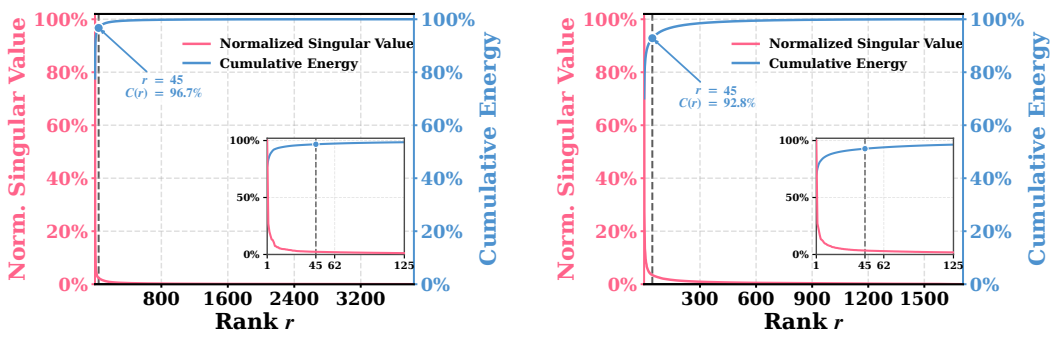
D Proof for Orthogonal Intervention

Proof of Theorem 1. For compactness, let $\mathbf{P} = \mathbf{P}_K$, $\mathbf{H} = \mathbf{H}_U$, $\mathbf{V} = \mathbf{V}_{\text{ref}}$, and $\mathbf{X} = \mathbf{PH}$. Consider first the objective before choosing a canonical representative. For an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, define its effective operator as

$$\mathbf{B} = \mathbf{AP}. \tag{26}$$

Because $\mathbf{P}^2 = \mathbf{P}$, we have $\mathbf{BP} = \mathbf{B}$. Furthermore,

$$\mathbf{APH} = \mathbf{BH} = \mathbf{BPH} = \mathbf{BX}, \qquad \|\mathbf{AP}\|_F = \|\mathbf{B}\|_F. \tag{27}$$



(a) EntityQA (b) PopQA

Figure 15: Singular value spectra and cumulative energy of known representations on Mistral-7B-Instruct-v0.3. The marker in each panel indicates $C(45)$.

Thus, the original objective uniquely identifies the effective operator through

$$\mathbf{B}_\alpha = \underset{\mathbf{BP}=\mathbf{B}}{\arg\min} \|\mathbf{BX} - \mathbf{V}\|_F^2 + \alpha \|\mathbf{B}\|_F^2 . \tag{28}$$

We first solve the corresponding unconstrained ridge problem. Define

$$\mathcal{L}_\alpha(\mathbf{B}) = \|\mathbf{BX} - \mathbf{V}\|_F^2 + \alpha \|\mathbf{B}\|_F^2 . \tag{29}$$

For every nonzero perturbation $\mathbf{\Delta} \in \mathbb{R}^{d\times d}$, its second directional derivative is

$$D^2\mathcal{L}_\alpha(\mathbf{B})[\mathbf{\Delta}, \mathbf{\Delta}] = 2\|\mathbf{\Delta X}\|_F^2 + 2\alpha \|\mathbf{\Delta}\|_F^2 > 0. \tag{30}$$

Hence, $\mathcal{L}_\alpha$ is strictly convex. Its unique stationary point satisfies

$$\nabla_{\mathbf{B}}\mathcal{L}_\alpha(\mathbf{B}) = 2(\mathbf{BX} - \mathbf{V})\mathbf{X}^\top + 2\alpha\mathbf{B} = \mathbf{0}, \tag{31}$$

$$\mathbf{B}\left(\mathbf{XX}^\top + \alpha\mathbf{I}_d\right) = \mathbf{VX}^\top. \tag{32}$$

Since $\alpha > 0$, $\mathbf{XX}^\top + \alpha\mathbf{I}_d$ is positive definite. The unique unconstrained minimizer is therefore

$$\mathbf{B}_\alpha = \mathbf{VX}^\top\left(\mathbf{XX}^\top + \alpha\mathbf{I}_d\right)^{-1}. \tag{33}$$

It remains to verify that Eq. (33) satisfies the constraint in Eq. (28). Since $\mathbf{X} = \mathbf{PH}$ and $\mathbf{P}^\top = \mathbf{P}$,

$$\mathbf{PX} = \mathbf{X}, \qquad \mathbf{X}^\top\mathbf{P} = \mathbf{X}^\top, \qquad \mathbf{PXX}^\top = \mathbf{XX}^\top\mathbf{P} = \mathbf{XX}^\top. \tag{34}$$

Consequently, $\mathbf{P}$ commutes with $\mathbf{XX}^\top + \alpha\mathbf{I}_d$ and its inverse. Using Eq. (34),

$$\mathbf{B}_\alpha\mathbf{P} = \mathbf{VX}^\top\left(\mathbf{XX}^\top + \alpha\mathbf{I}_d\right)^{-1}\mathbf{P} \tag{35}$$

$$= \mathbf{VX}^\top\mathbf{P}\left(\mathbf{XX}^\top + \alpha\mathbf{I}_d\right)^{-1} \tag{36}$$

$$= \mathbf{B}_\alpha. \tag{37}$$

Thus, the unconstrained minimizer is feasible for Eq. (28) and is also its unique minimizer.

Substituting $\mathbf{X} = \mathbf{P}_K\mathbf{H}_U$ and $\mathbf{V} = \mathbf{V}_{\text{ref}}$ into Eq. (33) gives

$$\mathbf{B}_\alpha = \mathbf{V}_{\text{ref}}\mathbf{H}_U^\top\mathbf{P}_K\left(\mathbf{P}_K\mathbf{H}_U\mathbf{H}_U^\top\mathbf{P}_K + \alpha\mathbf{I}_d\right)^{-1}. \tag{38}$$

All matrices $\mathbf{A}$ that induce this effective optimum satisfy $\mathbf{AP} = \mathbf{B}_\alpha$ and can be written as

$$\mathbf{A} = \mathbf{B}_\alpha + \mathbf{C}(\mathbf{I}_d - \mathbf{P}) \tag{39}$$

for an arbitrary $\mathbf{C} \in \mathbb{R}^{d\times d}$. The two summands in Eq. (39) are orthogonal in the Frobenius inner product because $\mathbf{B}_\alpha\mathbf{P} = \mathbf{B}_\alpha$. Therefore,

$$\|\mathbf{A}\|_F^2 = \|\mathbf{B}_\alpha\|_F^2 + \|\mathbf{C}(\mathbf{I}_d - \mathbf{P})\|_F^2 . \tag{40}$$

The unique minimum-Frobenius-norm representative is consequently $\mathbf{\Gamma}_\alpha = \mathbf{B}_\alpha$, which proves Eq. (10) and $\mathbf{\Gamma}_\alpha\mathbf{P}_K = \mathbf{\Gamma}_\alpha$. Finally, using $\mathbf{P}_K\mathbf{H}_K = \mathbf{0}$,

$$\mathbf{\Gamma}_\alpha\mathbf{H}_K = \mathbf{\Gamma}_\alpha\mathbf{P}_K\mathbf{H}_K = \mathbf{0}, \tag{41}$$

which proves Eq. (11) and completes the proof. □

E Additional Experimental Details

E.1 Dataset Preparation

We prepare question and answer records from TriviaQA, PopQA, and EntityQA (Joshi et al., 2017; Mallen et al., 2023; Ferrando et al., 2025). Each record has a unique identifier, a question, and one or more valid answers. We store the processed records in JSONL format.

TriviaQA. TriviaQA contains open domain trivia questions with answer aliases. We use the training partitions of the web and Wikipedia no context configurations. They contain 76,496 and 61,888 questions, respectively, for 138,384 records in total. We retain each question, its answer object and aliases, and its original question source. Each new identifier contains the dataset name, source partition, and original question identifier. We reject duplicate identifiers. Examples include:

- *Question:* From which country did Angola achieve independence in 1975? *Answer:* Portugal.
- *Question:* Who won Super Bowl XX? *Answer:* Chicago Bears.

PopQA. PopQA contains factual questions formed from subject, relation, and object triples. We use all 14,267 questions in the released TSV file. They cover 16 relation types, including occupation, place of birth, genre, director, author, country, and capital. We retain each question, its main object answer and valid aliases, its relation type, and its source Wikidata identifiers. We convert relation names to lowercase alphanumeric strings. Examples include:

- *Question:* What is the capital of Massac County? *Answer:* Metropolis. An accepted alias is Metropolis, Illinois.
- *Question:* Who was the director of Doomwatch? *Answer:* Peter Sasdy.

EntityQA. EntityQA contains questions about songs, movies, cities, and players. We use structured entity records from the released `sae_entities` data. We turn each entity and attribute pair into one question. The question templates cover 17 attribute types: four for songs, six for movies, four for cities, and three for players. This gives 153,239 questions: 33,792 about songs, 65,370 about movies, 31,616 about cities, and 22,461 about players. Examples include:

- *Question:* Who is the director of the movie *Stranger by the Lake*? *Answer:* Alain Guiraudie.
- *Question:* What country contains the city of McKinney? *Answer:* United States of America.

These totals describe the full prepared datasets. We label the questions separately for each model, as described in Section E.2, and exclude ambiguous questions. The tables in that subsection report the retained counts before experimental sampling. For each model, we use a fixed random seed to select 10,240 labeled questions from each dataset. We split each selection into training, validation, and test sets at an 80/10/10 ratio. This gives 8,192 training, 1,024 validation, and 1,024 test questions per dataset.

E.2 Knowledge-State Labeling

Knowledge state depends on the model. For each question, we sample ten responses from the target model. The prompt contains the question in one user message and uses the model's chat template. We sample with temperature 0.7, top-p 0.8, and top-k 20. Each response has at most 100 new tokens. We use a batch size of 8 and random seed 42.

We compare each response with the reference answer. For TriviaQA, valid answers include the normalized aliases and available human answers. For PopQA, they include all object aliases stored with the record. For both datasets, we lowercase the text, replace underscores with spaces, remove punctuation and the articles a, an, and the, and collapse extra spaces. A response is correct if it contains any normalized valid answer. For EntityQA, we lowercase the response and stored answer

and remove nonalphanumeric characters. A response is correct if it contains the normalized stored answer.

Let c_i be the number of correct responses among the ten samples for question Q_i. We assign the binary label y_i from Section 2 as

$$y_i = \begin{cases} 0, & c_i \geq 7, \\ 1, & c_i \leq 3. \end{cases} \tag{42}$$

Here, 0 denotes a known question and 1 denotes an unknown question. Questions with $4 \leq c_i \leq 6$ are ambiguous and do not enter the labeled dataset. Refusal detection uses a separate list of refusal phrases and does not affect the knowledge-state label.

Tables 4, 5, and 6 give the retained counts for each model and dataset. TriviaQA is grouped by source partition, EntityQA by entity type, and PopQA by relation type. The counts differ across models because each model receives its own labels. Across all three models, most retained TriviaQA questions are known, while most retained EntityQA and PopQA questions are unknown.

Table 4: Known and unknown question counts after labeling with LLaMA-3.1-8B-Instruct. Ambiguous questions are excluded.

Dataset	Subset	Known	Unknown	Total
TriviaQA	All Web	52,689	18,958	71,647
	All Wikipedia	45,723	12,468	58,191
	Total	**98,412**	**31,426**	**129,838**
EntityQA	City	4,360	26,619	30,979
	Movie	7,176	56,580	63,756
	Player	880	21,289	22,169
	Song	5,379	27,010	32,389
	Total	**17,795**	**131,498**	**149,293**
PopQA	Author	345	1,112	1,457
	Capital	511	116	627
	CapitalOf	221	121	342
	Color	26	5	31
	Composer	196	731	927
	Country	505	252	757
	Director	411	1,494	1,905
	Father	177	371	548
	Genre	345	1,149	1,494
	Mother	49	128	177
	Occupation	86	430	516
	PlaceOfBirth	33	541	574
	Producer	188	1,256	1,444
	Religion	80	215	295
	Screenwriter	348	1,570	1,918
	Sport	136	394	530
	Total	**3,657**	**9,885**	**13,542**

For illustration, we select one known and one unknown question from each dataset for LLaMA-3.1-8B-Instruct. Each selected known question has ten correct responses, and each selected unknown question has none.

- TriviaQA, known: From which country did Angola achieve independence in 1975? Reference answer: Portugal. Correct samples: $10/10$.
- TriviaQA, unknown: In which branch of the arts is Patricia Neary famous? Reference answer: Ballet. Correct samples: $0/10$.
- PopQA, known: What is Robby Krieger's occupation? Reference answer: guitarist. Correct samples: $10/10$.
- PopQA, unknown: What is George Rankin's occupation? Reference answer: politician. Correct samples: $0/10$.

Table 5: Known and unknown question counts after labeling with Qwen2.5-7B-Instruct. Ambiguous questions are excluded.

Dataset	Subset	Known	Unknown	Total
TriviaQA	All Web	49,031	23,511	72,542
	All Wikipedia	42,707	16,087	58,794
	Total	**91,738**	**39,598**	**131,336**
EntityQA	City	4,962	26,295	31,257
	Movie	2,944	61,361	64,305
	Player	324	21,880	22,204
	Song	1,736	31,148	32,884
	Total	**9,966**	**140,684**	**150,650**
PopQA	Author	189	1,295	1,484
	Capital	503	126	629
	CapitalOf	203	150	353
	Color	20	12	32
	Composer	118	836	954
	Country	608	170	778
	Director	122	1,850	1,972
	Father	162	381	543
	Genre	281	1,237	1,518
	Mother	41	135	176
	Occupation	139	346	485
	PlaceOfBirth	81	479	560
	Producer	106	1,367	1,473
	Religion	116	173	289
	Screenwriter	199	1,756	1,955
	Sport	247	262	509
	Total	**3,135**	**10,575**	**13,710**

Table 6: Known and unknown question counts after labeling with Mistral-7B-Instruct-v0.3. Ambiguous questions are excluded.

Dataset	Subset	Known	Unknown	Total
TriviaQA	All Web	54,894	18,196	73,090
	All Wikipedia	47,525	11,714	59,239
	Total	**102,419**	**29,910**	**132,329**
EntityQA	City	5,100	26,229	31,329
	Movie	8,077	55,677	63,754
	Player	1,090	20,980	22,070
	Song	4,636	27,963	32,599
	Total	**18,903**	**130,849**	**149,752**
PopQA	Author	349	1,120	1,469
	Capital	494	122	616
	CapitalOf	209	140	349
	Color	29	3	32
	Composer	197	735	932
	Country	647	132	779
	Director	403	1,530	1,933
	Father	191	348	539
	Genre	380	1,166	1,546
	Mother	59	112	171
	Occupation	166	331	497
	PlaceOfBirth	101	455	556
	Producer	267	1,179	1,446
	Religion	104	191	295
	Screenwriter	496	1,422	1,918
	Sport	267	245	512
	Total	**4,359**	**9,231**	**13,590**

- EntityQA, known: What is the album of the song *Burn It to the Ground*? Reference answer: Dark Horse. Correct samples: $10/10$.
- EntityQA, unknown: Who is the performer of the song *Special Needs*? Reference answer: Placebo. Correct samples: $0/10$.

E.3 LLM CONFIGURATIONS

We evaluate LLaMA-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3. We use the instruction tuned checkpoint of each model.

Common generation setup. We use the tokenizer and chat template provided with each model. We set the padding token to the end of sequence token and apply left padding. We load model weights in `bfloat16` and set the model to evaluation mode. Each question produces one sampled response. We use a batch size of 1 and generate at most 64 new tokens. We set the temperature to 1.0, top p to 0.9, and top k to 50. These decoding settings are shared by all methods. Each baseline changes only its prompt or the specified model activations.

Original LLM. We run each checkpoint without demonstrations or activation hooks. The input contains one test question in a single user message. We apply the model chat template and add its generation prompt. We then use the common generation setup above.

F IMPLEMENTATION DETAILS OF HALLUREFUSE AND BASELINES

All methods use the model and decoding settings described in Appendix E.3. This section provides implementation details for HALLUREFUSE and the baselines, including prompt construction, activation intervention, and method-specific settings.

HalluRefuse. We construct the steering matrix at decoder layer $l = 19$, following the layer indexing used in our implementation. We form $\mathbf{H}_K$ from the known steering, training, and validation examples. We form $\mathbf{H}_U$ from the refused and hallucinated steering examples, together with the unknown training and validation examples. We compute $\mathbf{V}_{\text{ref}}$ as the difference between the mean activations of refused and hallucinated unknown examples. We obtain $\mathbf{P}_K$ from the eigendecomposition of $\mathbf{H}_K^\top \mathbf{H}_K$. With the default ratio parameters, the code selects the numerical or structural null space and retains at least one direction. We then fit $\mathbf{\Gamma}_\alpha$ in this subspace with $\alpha = 10$ and set the inference coefficient λ to 1. The intervention is applied only during prompt encoding.

5-shot ICL. We build one fixed prompt from the training split. It contains one unknown example followed by four known examples. For the unknown example, we use the answer `I am unable to answer this question.` For each known example, we use the first model completion labeled as correct. We take the first required examples from the unknown and known training splits. Thus, the demonstration set is fixed and is not sampled for each query. We format every demonstration as a `Question:` block followed by an `Answer:` block. We append the test question with an empty answer block. The five demonstrations and the test question are placed in one user message.

ITI. ITI learns a refusal direction for each attention head (Li et al., 2023). We extract each head output at the input of the attention output projection. We use the final token of each chat formatted training prompt. An unknown example is positive when its refusal score is at least 0.5. It is negative when its refusal score is below 0.5 and its accuracy score is below 0.5. Correct unknown examples are not used to train the binary probes. We train one logistic regression probe for every layer and head. We use the training unknown split for fitting and the validation unknown split for head selection. Activation extraction uses a batch size of 8 and stores the tensors in `float16`. Each probe uses a random seed of 42 and at most 1000 iterations. We select the 48 heads with the highest validation accuracy.

For each selected head (l, h), we compute a unit direction $\mathbf{d}_{l,h}$. It is the normalized difference between the refusal and hallucination centroids from the combined training and validation data. We compute $\sigma_{l,h}$ as the population standard deviation of the training activations projected onto this

direction. At generation time, we edit the final token head output as

$$\mathbf{z}'_{l,h} = \mathbf{z}_{l,h} + \lambda \sigma_{l,h} \mathbf{d}_{l,h}. \tag{43}$$

We sweep $\lambda \in \{1, 2, 3, 4, 5, 6\}$. The selected heads and their directions are fixed before test generation.

CAA. CAA adds a fixed refusal direction to the residual stream (Rimsky et al., 2024). For each layer l, we load the cached direction

$$\mathbf{v}_{\text{ref}}^{(l)} = \overline{\mathbf{h}}_{\text{ref}}^{(l)} - \overline{\mathbf{h}}_{\text{hall}}^{(l)}, \tag{44}$$

where the two means come from refused and hallucinated unknown training questions. The cached CAA directions are not normalized. We cast them to `bfloat16` before generation. At a selected layer, we apply

$$\mathbf{h}^{(l)\prime} = \mathbf{h}^{(l)} + \lambda \mathbf{v}_{\text{ref}}^{(l)}. \tag{45}$$

Only selected layers have a nonzero strength. For LLaMA and Qwen2, we use the corresponding CAA steering model wrappers. For Mistral, we add the vector before the decoder layer input normalization during prompt prefill. In the output level study, we sweep $\lambda \in \{1, 2, 3, 4, 5, 6\}$. For LLaMA-3.1-8B-Instruct, we intervene at layer 20.

SADI-HEAD. SADI-HEAD selects attention heads from their activation differences (Wang et al., 2025). We use refused unknown training examples as the positive group. We use hallucinated unknown training examples as the negative group. For every layer and head, we first compute the difference between the two mean head output vectors. We then average this vector over its head dimension to obtain one scalar score. We rank all heads from all layers by this score and select the four largest entries. At generation time, we multiply the current output of each selected head by λ. The edit is applied to the final token at the input of the attention output projection. The setting $\lambda = 1$ leaves the activation unchanged. We use random seed 42 and sweep $\lambda \in \{1, 2, 3, 4, 5, 6\}$.

Spherical Steering. We follow the activation rotation in Spherical Steering (You et al., 2026). We load the same cached refusal direction as the target prototype $\boldsymbol{\mu}_T$. We normalize the prototype of each layer to unit length in `float32`. We set the opposite prototype to $\boldsymbol{\mu}_H = -\boldsymbol{\mu}_T$. We apply the spherical edit to the output of decoder layer 22. We fix $\kappa = 20$ and $\beta = 0.1$. We sweep $\alpha \in \{0.025, 0.05, 0.075, 0.1, 0.125, 0.15\}$.

G Generalization Evaluation Details

We evaluate transfer among TriviaQA, EntityQA, and PopQA. For each LLM, we hold out one dataset as the target and use the other two as sources. Each dataset serves as the target once. Knowledge states are defined separately for each LLM, as in the main experiments. The target dataset is used only for evaluation.

We construct a separate HalluRefuse intervention for each LLM and target choice. Known questions from the two source datasets form H_K. Unknown source questions form H_U, while refused and hallucinated unknown source questions provide the refusal direction. We compute P_K and Γ_α using the same procedure as in the main experiments, with $\alpha = 10$. At inference time, we apply the resulting matrix at decoder layer 19 with $\lambda = 1$. Every other layer has zero intervention strength.

We evaluate CAA, ITI, SADI-HEAD, and Spherical Steering under the same source and target rotation. Their intervention features come from the two source datasets. CAA and Spherical Steering use a refusal vector constructed from the unknown refusal and hallucination features of both sources. CAA applies it at layer 6 with strength 1. Spherical Steering uses normalized prototypes and applies its edit at layer 22, with $\alpha = 0.075$, $\kappa = 20$, and $\beta = 0.1$. ITI receives head output features from both sources. It selects 48 heads using probes, uses center of mass directions, and sets the strength to 1. SADI-HEAD also receives head output features from both sources. It selects the four heads with the largest scores and sets the strength to 5. These parameter settings are the same for all three LLMs.

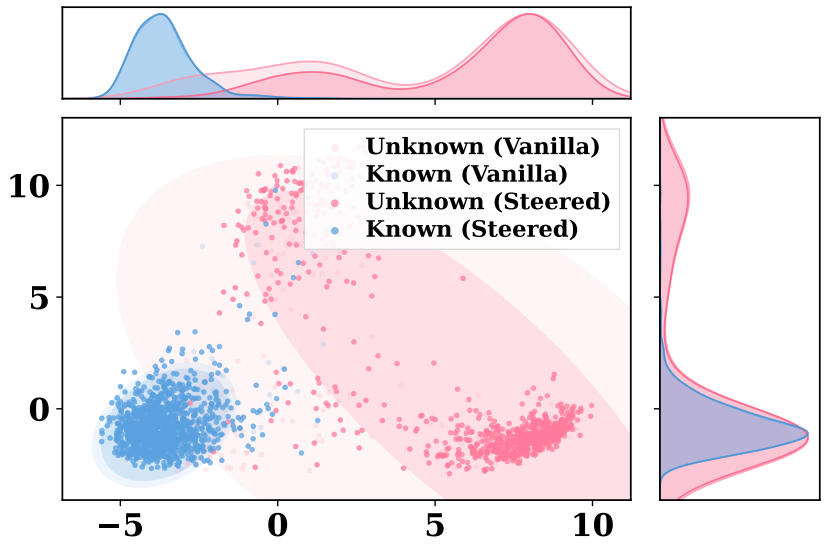


Figure 16: PCA view of known and unknown representations on TriviaQA before and after HalluRefuse steering. The known distributions nearly overlap.

For every method, we evaluate the known and unknown test splits of the target dataset. Sampling is enabled for all three LLMs. We use a batch size of 1, generate one response per question, and limit generation to 64 new tokens. Correctness and refusal are scored as in the main experiments. We report known accuracy, known refusal, and unknown hallucination for each target. The average columns in Table 2 are arithmetic means over the three targets for each LLM.

H PCA Views of Known Representations After HalluRefuse Steering

HalluRefuse protects the known subspace during steering. In particular, its intervention satisfies $\Gamma_\alpha H_K = 0$ for the protected known states. We use PCA to view the representations before and after steering. Figures 16, 17, and 18 show the results on LLaMA-3.1-8B-Instruct. Each plot includes known and unknown questions. Light shades show the original representations, and dark shades show the steered representations.

Across all three datasets, the dense known clusters remain close to their original positions. The marginal distributions also largely overlap. PopQA has a wider tail, but its main known cluster stays in place. Compared with the baseline PCA views in Appendix B.2, these plots show much less change in the known distributions. This pattern is consistent with the known subspace protection in Section 4.2.

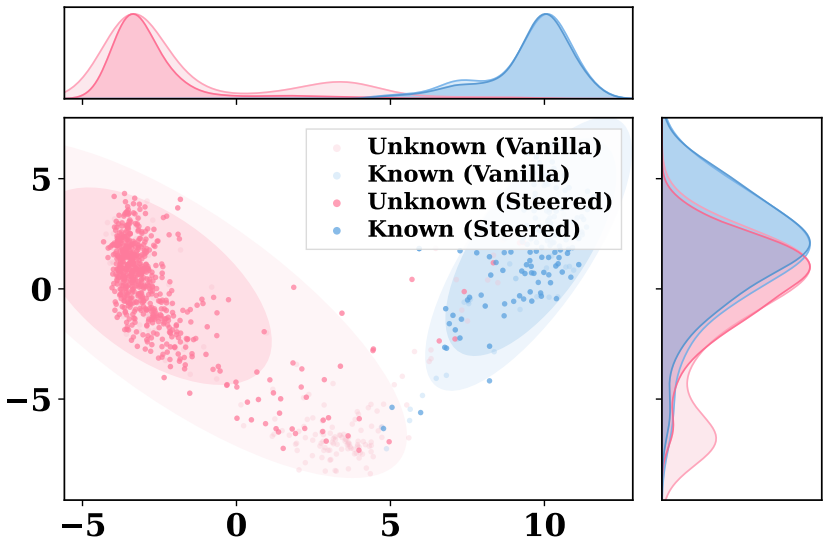


Figure 17: PCA view of known and unknown representations on EntityQA before and after HalluRefuse steering. The main known cluster stays in the same region.

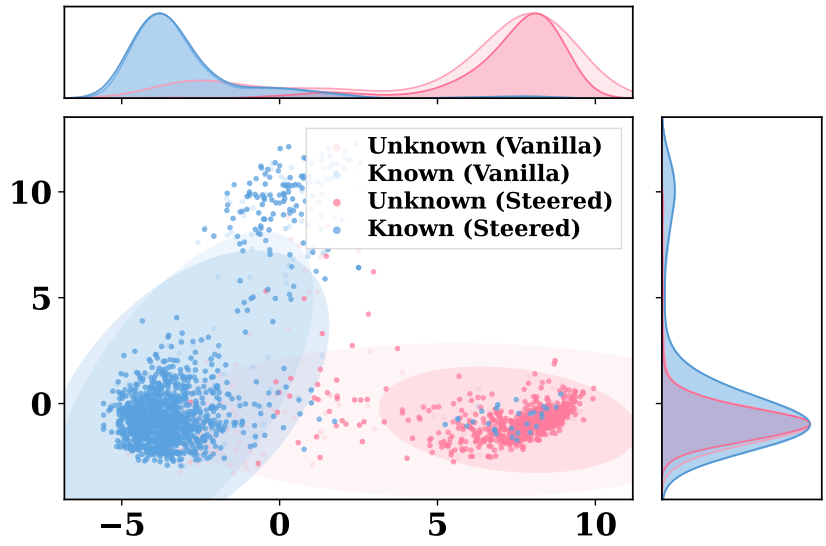


Figure 18: PCA view of known and unknown representations on PopQA before and after HalluRefuse steering. The main known cluster remains close to its original position.

I Example Mode Outputs

We show ten generations for each of four unknown and four known questions from LLaMA-3.1-8B-Instruct after applying HALLUREFUSE. For the unknown questions, responses say that the model cannot verify the answer. For the known questions, the responses retain the main answer. These examples show the pattern of refusal on unknown questions and answer preservation on known questions, along with variation in the generated text.

After HALLUREFUSE Steering: Example Unknown Question #1

Query: *What is the name of Hyacinth and Richard's son in 'Keeping Up Appearances'?*

LLaMA-3.1-8B-Instruct; Generation #1: I am unable to verify the name of Hyacinth and Richard's son in the TV show 'Keeping Up Appearances'.

LLaMA-3.1-8B-Instruct; Generation #2: I am sorry, I cannot verify the name of Hyacinth and Richard's son in 'Keeping Up Appearances'.

LLaMA-3.1-8B-Instruct; Generation #3: I am unable to verify the name of Hyacinth and Richard's son in the TV show 'Keeping Up Appearances'.

LLaMA-3.1-8B-Instruct; Generation #4: I'm not aware of any information about a character named Hyacinth and Richard's son in the TV series 'Keeping Up Appearances'.

LLaMA-3.1-8B-Instruct; Generation #5: I am unable to verify the name of Hyacinth and Richard's son in the TV series 'Keeping Up Appearances'.

LLaMA-3.1-8B-Instruct; Generation #6: I don't have information about the name of Hyacinth and Richard's son in the TV show 'Keeping Up Appearances'.

LLaMA-3.1-8B-Instruct; Generation #7: I couldn't find information about the name of Hyacinth and Richard's son in the TV show 'Keeping Up Appearances'.

LLaMA-3.1-8B-Instruct; Generation #8: I couldn't find any information about the name of Hyacinth and Richard's son in the TV show 'Keeping Up Appearances'.

LLaMA-3.1-8B-Instruct; Generation #9: I am unable to verify the religion of Father

LLaMA-3.1-8B-Instruct; Generation #10: I couldn't find any information about a character named Hyacinth and Richard's son in the TV series 'Keeping Up Appearances'.

After HALLUREFUSE Steering: Example Unknown Question #2

Query: *Upton Park Football Club won what notable international award in Paris, 1900?*

LLaMA-3.1-8B-Instruct; Generation #1: I cannot verify what award Upton Park Football Club won in Paris, 1900.

LLaMA-3.1-8B-Instruct; Generation #2: I am unable to verify what notable international award Upton Park Football Club won in Paris in 1900.

LLaMA-3.1-8B-Instruct; Generation #3: I am unable to verify what notable international award Upton Park Football Club won in Paris, 1900.

LLaMA-3.1-8B-Instruct; Generation #4: I am unable to verify what international award Upton Park Football Club won in Paris in 1900.

LLaMA-3.1-8B-Instruct; Generation #5: I am unable to verify what international award Upton Park Football Club won in Paris in 1900.

LLaMA-3.1-8B-Instruct; Generation #6: I am unable to verify what award Upton Park Football Club won in Paris in 1900.

LLaMA-3.1-8B-Instruct; Generation #7: I am unable to verify any information about Upton Park Football Club winning a notable international award in Paris in 1900.

LLaMA-3.1-8B-Instruct; Generation #8: I cannot verify the information on Upton Park Football Club winning an international award in Paris in 1900.

LLaMA-3.1-8B-Instruct; Generation #9: I am unable to verify Upton Park Football Club winning a notable international award in Paris in 1900.

LLaMA-3.1-8B-Instruct; Generation #10: I cannot verify what award Upton Park Football Club won in Paris in 1900.

After HALLUREFUSE Steering: Example Unknown Question #3

Query: *Which British monarch donated the King's Cup for the Cowes Week Race?*

LLaMA-3.1-8B-Instruct; Generation #1: I am unable to verify who donated the King's Cup for the Cowes Week Race.

LLaMA-3.1-8B-Instruct; Generation #2: I am unable to verify which British monarch donated the King's Cup for the Cowes Week Race.

LLaMA-3.1-8B-Instruct; Generation #3: I am unable to verify who donated the King's Cup for the Cowes Week Race.

LLaMA-3.1-8B-Instruct; Generation #4: I am unable to verify which British monarch donated the King's Cup for the Cowes Week Race.

LLaMA-3.1-8B-Instruct; Generation #5: I am unable to verify which British monarch donated the King's Cup for the Cowes Week Race.

LLaMA-3.1-8B-Instruct; Generation #6: I cannot verify who donated the King's Cup for the Cowes Week Race.

LLaMA-3.1-8B-Instruct; Generation #7: I am unable to verify any information about who donated the King's Cup for the Cowes Week Race.

LLaMA-3.1-8B-Instruct; Generation #8: I am unable to verify who donated the King's Cup for the Cowes Week Race.

LLaMA-3.1-8B-Instruct; Generation #9: I am unable to verify which British monarch donated the King's Cup for the Cowes Week Race.

LLaMA-3.1-8B-Instruct; Generation #10: I am unable to verify which British monarch donated the King's Cup for the Cowes Week Race.

After HALLUREFUSE Steering: Example Unknown Question #4

Query: *In 1969, P H Newby's book Something to Answer For was the first winner of what prize?*

LLaMA-3.1-8B-Instruct; Generation #1: I do not have information on P H Newby's book Something to Answer For being the first winner of a prize in 1969.

LLaMA-3.1-8B-Instruct; Generation #2: I am unable to verify the winner of the 1969 prize.

LLaMA-3.1-8B-Instruct; Generation #3: I am unable to verify what the first winner of the 1969 Booker Prize was.

LLaMA-3.1-8B-Instruct; Generation #4: I am unable to verify the winner of the prize in 1969 for the book Something to Answer For.

LLaMA-3.1-8B-Instruct; Generation #5: I am unable to verify what book was the first winner of the Booker Prize in 1969

LLaMA-3.1-8B-Instruct; Generation #6: I cannot verify the first winner of the prize in 1969 for P H Newby's book "Something to Answer For".

LLaMA-3.1-8B-Instruct; Generation #7: I cannot verify the winner of the 1969 prize.

LLaMA-3.1-8B-Instruct; Generation #8: I am unable to verify the information about P H Newby's book Something to Answer For.

LLaMA-3.1-8B-Instruct; Generation #9: I cannot verify what prize P H Newby's book "Something to Answer For" won in 1969.

LLaMA-3.1-8B-Instruct; Generation #10: I am unable to verify what prize Something to Answer For was the first winner of.

After HALLUREFUSE Steering: Example Known Question #1

Query: *Mark Almond was the lead singer of 80's pop duo Soft Cell, but who was the other member of the band?*

LLaMA-3.1-8B-Instruct; Generation #1: The other member of the 80's pop duo Soft Cell was Dave Ball.

LLaMA-3.1-8B-Instruct; Generation #2: The other member of the 80's pop duo Soft Cell was David Ball.

LLaMA-3.1-8B-Instruct; Generation #3: The other member of Soft Cell was David Ball.

LLaMA-3.1-8B-Instruct; Generation #4: The other member of the 80's pop duo Soft Cell was David Ball.

LLaMA-3.1-8B-Instruct; Generation #5: The other member of Soft Cell was Dave Ball.

LLaMA-3.1-8B-Instruct; Generation #6: The other member of the 80's pop duo Soft Cell was David Ball.

LLaMA-3.1-8B-Instruct; Generation #7: The other member of Soft Cell was David Ball.

LLaMA-3.1-8B-Instruct; Generation #8: The other member of the 80's pop duo Soft Cell was David Ball.

LLaMA-3.1-8B-Instruct; Generation #9: The other member of the 80's pop duo Soft Cell was David Ball.

LLaMA-3.1-8B-Instruct; Generation #10: The other member of the 80's pop duo Soft Cell was David Ball.

After HALLUREFUSE Steering: Example Known Question #2

Query: *What is the name of the process by which plants obtain carbohydrates from carbon dioxide and water?*

LLaMA-3.1-8B-Instruct; Generation #1: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

LLaMA-3.1-8B-Instruct; Generation #2: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

LLaMA-3.1-8B-Instruct; Generation #3: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

LLaMA-3.1-8B-Instruct; Generation #4: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

LLaMA-3.1-8B-Instruct; Generation #5: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

LLaMA-3.1-8B-Instruct; Generation #6: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

LLaMA-3.1-8B-Instruct; Generation #7: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

LLaMA-3.1-8B-Instruct; Generation #8: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

LLaMA-3.1-8B-Instruct; Generation #9: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

LLaMA-3.1-8B-Instruct; Generation #10: The process by which plants obtain carbohydrates from carbon dioxide and water is called photosynthesis.

After HALLUREFUSE Steering: Example Known Question #3

Query: *Which animal has species called Plainss, Mountain and Grevy's*

LLaMA-3.1-8B-Instruct; Generation #1: The animal you're referring to is the zebra.

LLaMA-3.1-8B-Instruct; Generation #2: The animal species you're referring to is the zebra.

LLaMA-3.1-8B-Instruct; Generation #3: The animal you are referring to is the zebra.

LLaMA-3.1-8B-Instruct; Generation #4: The animal with species called Plains, Mountain, and Grevy's is the zebra.

LLaMA-3.1-8B-Instruct; Generation #5: The animal with species called Plains, Mountain, and Grevy's is the Zebra.

LLaMA-3.1-8B-Instruct; Generation #6: The animal you are referring to is the Zebra.

LLaMA-3.1-8B-Instruct; Generation #7: The animal you are referring to is the Zebra.

LLaMA-3.1-8B-Instruct; Generation #8: The animal you are referring to is the Zebra.

LLaMA-3.1-8B-Instruct; Generation #9: The animal you are referring to is the Zebra.

LLaMA-3.1-8B-Instruct; Generation #10: The animal with species called Plains, Mountain, and Grevy's is the Zebra.

After HALLUREFUSE Steering: Example Known Question #4

Query: *In the Second World War, what was Operation Avalanche?*

LLaMA-3.1-8B-Instruct; Generation #1: Operation Avalanche was the codename for the Allied invasion of Italy, specifically the invasion of the Salerno beachhead, which occurred on September 9, 1943. It was a major military operation during World War II, where Allied forces landed on the southern coast of Italy.

LLaMA-3.1-8B-Instruct; Generation #2: Operation Avalanche was an Allied invasion of Italy during World War II. It began on September 3, 1943, and involved the Allied invasion of the Italian mainland, specifically the Salerno beachhead. The operation was carried out by the US Fifth Army, led by General Mark W. Clark, with the primary objective of capturing the port city of Naples and then moving on to Rome.

LLaMA-3.1-8B-Instruct; Generation #3: Operation Avalanche was an Allied invasion of Italy during World War II. It began on September 3, 1943, and was carried out by the Allies, primarily consisting of American and Canadian forces. The operation aimed to capture the Italian mainland, specifically the southern region of Italy, and was a crucial part of the Italian Campaign.

LLaMA-3.1-8B-Instruct; Generation #4: Operation Avalanche was the Allied invasion of Italy during World War II. It began on September 3, 1943, when Allied forces landed on the Italian mainland near the town of Reggio Calabria.

LLaMA-3.1-8B-Instruct; Generation #5: Operation Avalanche was the codename for the Allied invasion of Italy during World War II. It began on September 3, 1943, with the Allied invasion of the Italian mainland at Salerno, near the town of Paestum, which is about 25 miles (40 km) southeast of Naples. The operation was a key part of the Allied strategy to capture Italy and ultimately push the Axis powers out of the country.

LLaMA-3.1-8B-Instruct; Generation #6: Operation Avalanche was an Allied invasion of Italy during World War II. It was a major military operation conducted by the Allied forces, specifically the U.S. Fifth Army, led by General Mark W. Clark. The operation took place on September 3, 1943, and its main objective was to capture the city of Salerno, which is located in the southern part of Italy, near the Gulf of Salerno.

LLaMA-3.1-8B-Instruct; Generation #7: Operation Avalanche was an Allied military operation during World War II, specifically the Allied invasion of Italy. It was a major amphibious assault that took place on September 3, 1943, as part of the Allied invasion of Italy, which was code-named 'Operation Baytown' or 'Operation Avalanche'.

LLaMA-3.1-8B-Instruct; Generation #8: Operation Avalanche was an Allied invasion of Italy during World War II. It took place on September 3, 1943. The operation involved a large-scale amphibious assault by the U.S. Fifth Army, which landed on the southern coast of Italy, specifically at Salerno, near the town of Paestum.

LLaMA-3.1-8B-Instruct; Generation #9: Operation Avalanche was an Allied military operation during World War II, specifically the Allied invasion of Italy. It was a strategic invasion of the Italian mainland, specifically targeting the southern region of Italy.

LLaMA-3.1-8B-Instruct; Generation #10: Operation Avalanche was the codename for the Allied invasion of Italy, specifically the invasion of the Salerno beachhead, which occurred on September 9, 1943. The operation was launched by the Allied forces in the Mediterranean Theater, primarily consisting of American and British troops.