Detecting (Un)answerability in Large Language Models with Linear Directions

Anonymous ACL submission

Abstract

Large language models (LLMs) often respond confidently to questions even when they lack the necessary information, leading to hallucinated answers. In this work, we study the problem of (un)answerability detection in extractive question answering (QA), where the model should determine if a passage contains sufficient information to answer a given question. We propose a simple approach that identifies a direction in the model's activation space that captures unanswerability and uses it for classification. This direction is selected by applying activation additions during inference and measuring their impact on the model's abstention behavior. We show that projecting hidden activations onto this direction yields a reliable score for (un)answerability classification. Experiments on two open-weight LLMs and four QA benchmarks show that our method effectively detects unanswerable questions and generalizes better across datasets than existing prompt-based and classifier-based approaches. Causal interventions reveal that adding the direction increases abstention, while ablating it suppresses it, further indicating that it captures an unanswerability signal.[1]

1 Introduction

Large language models (LLMs) often generate confident responses to questions regardless of whether they have the information needed to answer reliably (Yin et al., 2023; Yona et al., 2024). When a model lacks the required information, it often produces inaccurate responses or hallucinations (Huang et al., 2025; Luo et al., 2024), making the identification of such cases an important step toward improving its trustworthiness (Kadavath et al., 2022; Yin et al., 2023; Amayuelas et al., 2024). This challenge is particularly important in applications such as medical assistance, legal advice, and educational tools, where incorrect answers can lead to real-world harm.

In this work, we study the problem of unanswerability in the context of extractive question answering (QA), where the model is presented with a question and a passage of text that may or may not contain the information required to answer it (Rajpurkar et al., 2018). As illustrated in Figure 1, models in this setting tend to respond rather than abstain, even when the question cannot be answered from the provided passage.

Several approaches have been proposed for detecting unanswerable questions. Fine-tuning has been suggested to improve abstention behavior in models (Feng et al., 2024; Zhang et al., 2024). In extractive QA, prompting has been shown to encourage models to indicate uncertainty (Slobodkin et al., 2023), but performance remains inconsistent across models and datasets. Slobodkin et al. (2023) further introduced a linear classifier trained on internal model representations to predict unanswerability. Other efforts have explored estimating uncertainty from hidden states (Tomani et al., 2024; Kim et al., 2024), or detecting unanswerable inputs with sparse autoencoder features (Heindrich et al., 2025). While these latter methods have shown promising results, they often fail to generalize across datasets—highlighting a key challenge in robust unanswerability detection.

Here, we analyze the model's internal activations and show that a single direction in representation space effectively captures unanswerability across diverse datasets. To this end, we first construct a set of candidate directions using difference-in-means (Marks and Tegmark, 2024), where the averaged activations of answerable examples are subtracted from those of unanswerable ones at a fixed layer and position. To select the most informative direction, we add each candidate vector to the hidden

[1] Our code is available at `https://anonymized`.

activations at inference time and measure the resulting change in the model's probability of abstaining. Finally, we use the selected direction for unanswerability classification: given an input, we extract its activations at a fixed layer and position and project it onto the learned direction. This projection yields a scalar unanswerability score, which reflects how aligned the model's internal representation is with unanswerable examples.

We evaluate our method on four question-answering datasets: SQuAD 2.0 (Rajpurkar et al., 2016, 2018), RepLiQA (Monteiro et al., 2024), Natural Questions (NQ) (Kwiatkowski et al., 2019), and MuSiQue (Trivedi et al., 2022), using Llama-3-8B-Instruct (Dubey et al., 2024) and Gemma-3-12B-IT (Team et al., 2025), and find that the learned direction consistently captures unanswerability. Our method achieves F1 scores of 75.9–96.4%, performing comparably to a logistic regression classifier baseline and outperforming prompt-based baselines. We also show that the direction signal transfers across datasets, exceeding the classifier's generalization on three out of four datasets by an average of 8.14%. Moreover, a simple threshold calibration using the validation split of each evaluation dataset further improves performance by 9.73% on average. These results highlight the robustness of the learned direction and its ability to generalize across datasets. We further validate the signal encoded by the direction through causal interventions, where adding the direction vector to the residual stream at a sufficient magnitude causes the model to abstain in nearly all cases (96%), while ablating it pushes the model to answer even when the context is insufficient.

Beyond classification, our method provides insight into how unanswerability is internally represented by the model, revealing a native signal embedded directly in the representation space. Analysis of failure cases further supports the reliability of this signal. In several instances (26%), we found that the provided labels were incorrect. Also, in 24% of cases labeled as answerable, the answer appeared in the passage but not in the context of the specific question, making the instance difficult to classify. A smaller portion (6%) included questions with grammatical issues, rendering their answerability unclear and dependent on interpretation.

To conclude, we introduce a lightweight and interpretable method for detecting unanswerability in LLMs by uncovering a direction in the model's activation space that captures an unanswerability signal. We demonstrate the utility of this approach for classifying unanswerable inputs across diverse datasets, and show that the learned direction generalizes better than prompt-based and classifier-based baselines. We also show that we can use this direction to control the model's tendency to abstain.

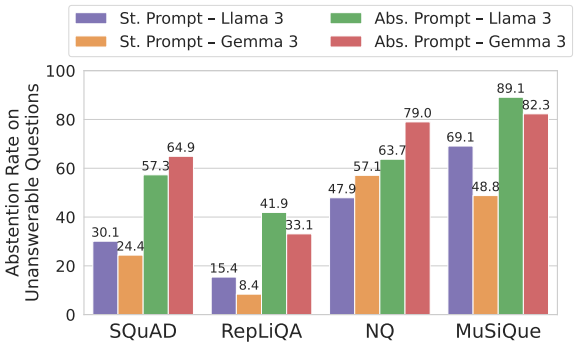

Figure 1: Abstention rate (recall) on unanswerable questions under Standard and Abstain-aware prompts, evaluated on Llama-3-8B-Instruct and Gemma-3-12B-IT.

2 Problem Setup

We address the task of *unanswerability detection* in extractive QA. Given a context (e.g., passage or document) c and a question q, the goal is to determine whether the context contains sufficient information to answer the question. Formally, the input is a pair (c, q), and the objective is to predict a binary label $y \in \{0, 1\}$, where $y = 1$ indicates that the question is unanswerable based on the context, and $y = 0$ indicates that it is answerable. Examples of answerable and unanswerable cases are shown in Table 1.

3 Method

We take inspiration from prior observations that certain abstract concepts, such as sentiment, refusal, or truthfulness, are linearly encoded within a language model's internal representations (Tigges et al., 2023; Arditi et al., 2024; Marks and Tegmark, 2024, inter alia), and aim to identify a direction in the model's activation space that captures unanswerability. If such a direction exists, it can be used to distinguish answerable from unanswerable instances by measuring the alignment between their internal representations and this direction. We now describe our methodology for finding such directions in LLMs.

Dataset	Context (c)	Question (q)	Label (y)
SQuAD	In England, the period of Norman architecture immediately succeeds that of the Anglo-Saxon and precedes the Early Gothic...	What architecture type came after Early Gothic?	1 (unanswerable)
RepLiQA	...One such partnership was formed with the Greenleaf Cafe, a popular downtown eatery, which now organizes 'Saturday Morning Miles'...	What specific event does the Greenleaf Cafe organize as part of Newville's fitness initiative?	0 (answerable)
NQ	The National Professional Soccer League II , which awarded two points for all goals except those on the power play , also used a three - point line...	when did the nba add the three point line ?	1 (unanswerable)
MuSiQue	Ye Rongguang (born October 3, 1963 in Wenzhou, Zhejiang) . . . Sanjiang Church was a Christian church located in Yongjia County, near Wenzhou, in Zhejiang Province, China...	What county was Ye Rongguang born in?	0 (answerable)

Table 1: Example context–question pairs from each dataset used in our experiments, labeled as answerable (0) or unanswerable (1).

3.1 Deriving Potential Directions

To identify potential directions encoding unanswerability, we follow prior work that uses differences in mean activations between two input sets (Marks and Tegmark, 2024; Belrose, 2024; Rimsky et al., 2024). Given a model with L layers and hidden dimension d, for each input (c, q) we extract the hidden activations $\mathbf{h}_{\ell,p} \in \mathbb{R}^d$ at each layer $\ell \in \{1, \ldots, L\}$ and token position p after the instruction segment.[2] Let $\{(c_i, q_i)\}_{i=1}^N$ be answerable and $\{(c_j, q_j)\}_{j=1}^M$ unanswerable examples, and let $\mathbf{h}_{\ell,p}^{(i)}$ be the hidden activations for the i-th input. We define the candidate direction $\mathbf{v}_{\ell,p} \in \mathbb{R}^d$ at each layer ℓ and token position p as the difference between the mean activations over unanswerable and answerable examples:

$$\mathbf{v}_{\ell,p} = \frac{1}{M}\sum_{j=1}^{M} \mathbf{h}_{\ell,p}^{(j)} - \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_{\ell,p}^{(i)}$$

This yields a set of $L \times N_{\text{pos}}$ directions $\{\mathbf{v}_{\ell,p}\}$, where N_{pos} is the number of token positions considered.

3.2 Selecting a Direction for Unanswerability

We employ causal steering (Li et al., 2023; Turner et al., 2023; Rimsky et al., 2024) to choose the direction that best represents unanswerability. The selection is done on a separate validation set from the examples used to find the candidate directions.

[2]These positions correspond to tokens from a chat template that wraps chat models' inputs and appear before the model's response, see §4.1 for details

Activation intervention For each candidate direction $\mathbf{v}_{\ell,p}$ and context-question pair (c, q) in the validation set, we modify the hidden activations at the corresponding layer ℓ and position p as follows:

$$\tilde{\mathbf{h}}_{\ell,p} = \mathbf{h}_{\ell,p} + \mathbf{v}_{\ell,p}.$$

The modified activations are propagated forward through the model. We repeat this procedure for each candidate direction and analyze its effect on the model's outputs and abstention behavior.

Steering score Let $\{(c_i, q_i)\}_{i=1}^K$ denote the validation set, consisting of K context-question pairs. To approximate abstention behavior, we identify the first token of the word *unanswerable* as it is tokenized by the model (e.g., "un"), and denote it as $t_{\text{un}} \in \mathcal{V}$. This token is used as a proxy for abstention since the model is prompted to respond with the word *unanswerable* when it cannot answer the question based on the provided context.

For each validation example, we extract the model's next-token distribution under the intervention. Let $p_t^{(i)}$ denote the probability of token t for the i-th validation example, the steering score ψ_{steer} of a direction $\mathbf{v}_{\ell,p}$ is then defined as:

$$\psi_{\text{steer}} = \frac{1}{K}\sum_{i=1}^{K}\left[\log p_{t_{\text{un}}}^{(i)} - \log \sum_{t \in \mathcal{V}\setminus\{t_{\text{un}}\}} p_t^{(i)}\right]$$

This score quantifies how much more likely the model is, on average, to generate t_{un} rather than any other token in the vocabulary, when steered with the candidate direction. Higher values indicate a stronger abstention-inducing effect.

Direction selection We evaluate all $L \times N_{\text{pos}}$ candidate directions and select the one with the highest steering score. The final unanswerability direction, denoted $\mathbf{v}^*$, corresponds to the pair (ℓ^*, p^*) that maximizes ψ_{steer}. This selected direction is used in all downstream evaluations and analyses.

3.3 Unanswerability Classification

We use the selected direction $\mathbf{v}^*$ to define a scalar scoring function that quantifies how strongly a given input aligns with the unanswerability signal; this score is then used to classify new inputs as answerable or unanswerable.

Unanswerability score Let $\hat{\mathbf{v}}^*$ denote the normalized direction selected in the previous step. For a given context-question pair (c, q), we extract the hidden activations $\mathbf{h}^* \in \mathbb{R}^d$ from the selected layer ℓ^* and position p^*. The unanswerability score is computed as the dot product between this hidden state and the normalized direction:

$$\phi_{\text{unans}} = \langle \mathbf{h}^*, \hat{\mathbf{v}}^* \rangle$$

This scalar is intended to reflect how strongly the input aligns with the learned unanswerability signal. Since it is unbounded and varies across models and datasets, we next describe how we interpret this value for classification.

Thresholding the unanswerability score To establish a classifier, we select the threshold τ on the unanswerability score using the validation set. Specifically, we compute the ROC curve and choose τ to minimize the Euclidean distance to the ideal point (TPR = 1, FPR = 0). At inference time, for a given input (c, q), if ϕ_{unans} exceeds τ, then the input is classified as unanswerable; otherwise, it is classified as answerable.

4 Experiments

We evaluate our method against three baselines and report classification accuracy and generalization across datasets, as well as a causal analysis of the learned direction. Our results show that: (1) the direction-based method achieves strong performance when derived and evaluated on the same dataset, close to a trained classifier and outperforming prompt-based baselines; (2) the direction generalizes more robustly across datasets than the classifier, especially after a lightweight threshold calibration; and (3) the selected directions causally influence the model's abstention behavior.

4.1 Experimental Setup

Datasets We evaluate our method on four question answering benchmarks—SQuAD 2.0 (Rajpurkar et al., 2018), RepLiQA (Monteiro et al., 2024), NQ (Kwiatkowski et al., 2019), and MuSiQue (Trivedi et al., 2022)—all structured as context-question pairs.

SQuAD 2.0 and RepLiQA natively include explicitly labeled answerable and unanswerable examples. SQuAD 2.0 augments the original SQuAD dataset, which is based on Wikipedia articles, with unanswerable questions that appear plausible given the context. RepLiQA is constructed from human-written reference documents across diverse topics not found on the web, so models cannot rely on their parametric knowledge.

For NQ and MuSiQue, we use the versions of the datasets curated by Slobodkin et al. (2023). NQ consists of real user questions paired with Wikipedia paragraphs, while MuSiQue contains multi-hop questions created by composing seed questions from various datasets. The curated version retains the original answerable examples and constructs unanswerable ones by replacing gold paragraphs with semantically similar ones that do not answer the question. (see §A for more details).

For each dataset, we sample a total of 4,000 examples, which we split into training (1,200), development (800), and test (2,000) sets, with an equal number of answerable and unanswerable instances in each split. Table 1 provides representative examples from each dataset.

Models We experiment with two instruction-tuned models: Llama-3-8B-Instruct (Dubey et al., 2024) and Gemma-3-12B-IT (Team et al., 2025). Both were trained with chat templates that wrap the user instruction (see §A.2 for the full templates). In our analysis, we focus on hidden activations at the positions of the template tokens that immediately follow the user instruction, as they represent the model's internal state after processing the full context and question and just before it begins generating a response. In addition, all inputs are formatted using the Abstain-aware Prompt (see §4.1).

Method We find that the token "un" corresponds to the first token in *unanswerable* in both Llama-3-8B-Instruct and Gemma-3-12B-IT, and set it as t_{un}. We apply the method described in §3.2 to select the layer and token position for each model–dataset pair, and find that the selected layers consistently

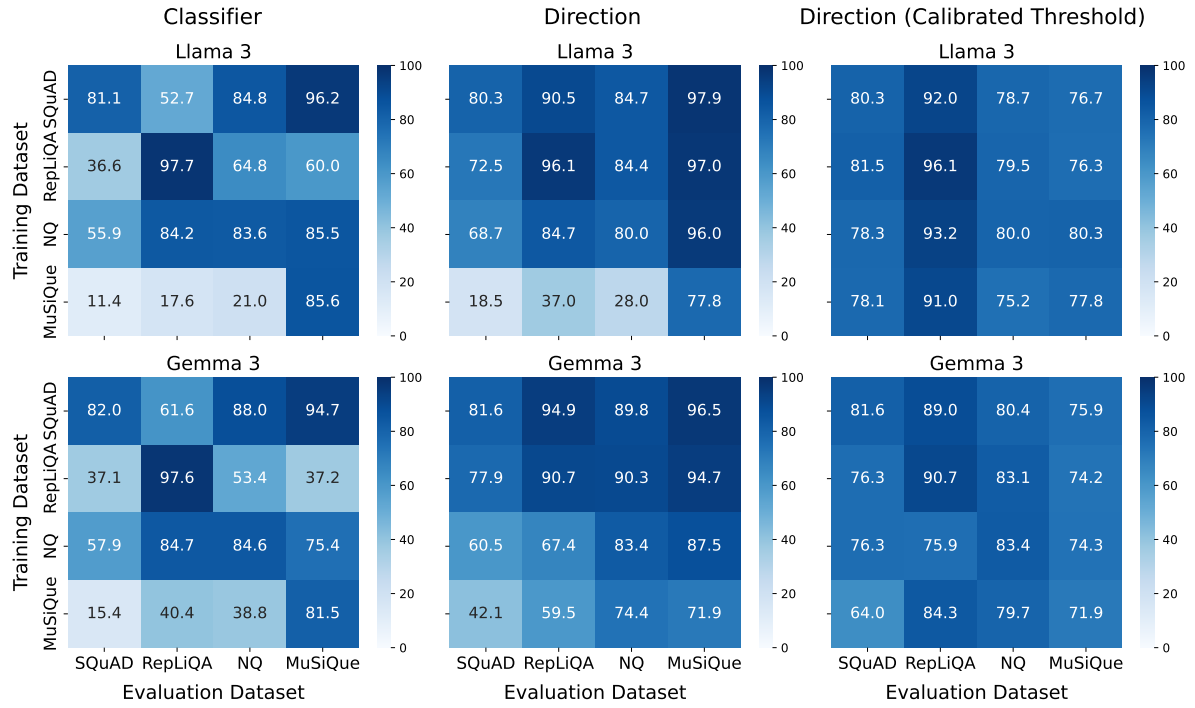

Figure 2: Unanswerable prompts recall (abstention rate) across datasets using three methods: a trained classifier, a direction-based method with a fixed threshold, and a calibrated threshold variant. Each heatmap shows generalization performance from training on one dataset (rows) to evaluating on another (columns). Results are shown for both Llama-3-8B-Instruct (top) and Gemma-3-12B-IT (bottom).

lie near the middle of the model. This is consistent with prior work suggesting that middle layers in transformer models tend to capture abstract semantic properties, in contrast to lower layers which focus on lexical patterns and upper layers which are more task-specific (Geva et al., 2021; Vulić et al., 2020; Tenney et al., 2019; Jawahar et al., 2019). Classification thresholds are set using ROC curves on the validation sets (see §A.5 for direction and threshold selection details).

Baselines We compare our method against the following baselines:

- *Standard Prompt*: A prompt-only baseline where the model is given the context and question without any additional instruction.
- *Abstain-aware Prompt*: A prompt augmentation baseline, in which an instruction is added encouraging the model to abstain if the question is unanswerable (Slobodkin et al., 2023).
- *Classifier*: A logistic regression model trained on hidden activations to predict unanswerability (Slobodkin et al., 2023). The classifier is trained using cross-validation on the combined training and validation sets, with model inputs formatted using the Abstain-aware Prompt.

Full prompt templates for the prompt-based baselines are provided in §A.3.

Evaluation metrics We measure precision, recall, and F1 score separately for the answerable and unanswerable classes. We also report macro-average F1 score, which balances precision and recall across both classes equally. Since the prompt-based baselines generate textual output, we first classify each response as either an abstention or an attempt to answer the question. To do so, we use GPT-4o mini (OpenAI, 2024), prompted with instructions and few-shot examples to make this decision. The full prompt used and a manual analysis validating this automatic evaluation are in §A.4.

4.2 (Un)answerability Classification

We evaluate how effectively our method distinguishes answerable and unanswerable questions.

Direction-based method effectively detects unanswerable questions Figure 2 (left and middle) shows the recall on unanswerable examples for our direction-based method and the classifier baseline, across all combinations of evaluation and source

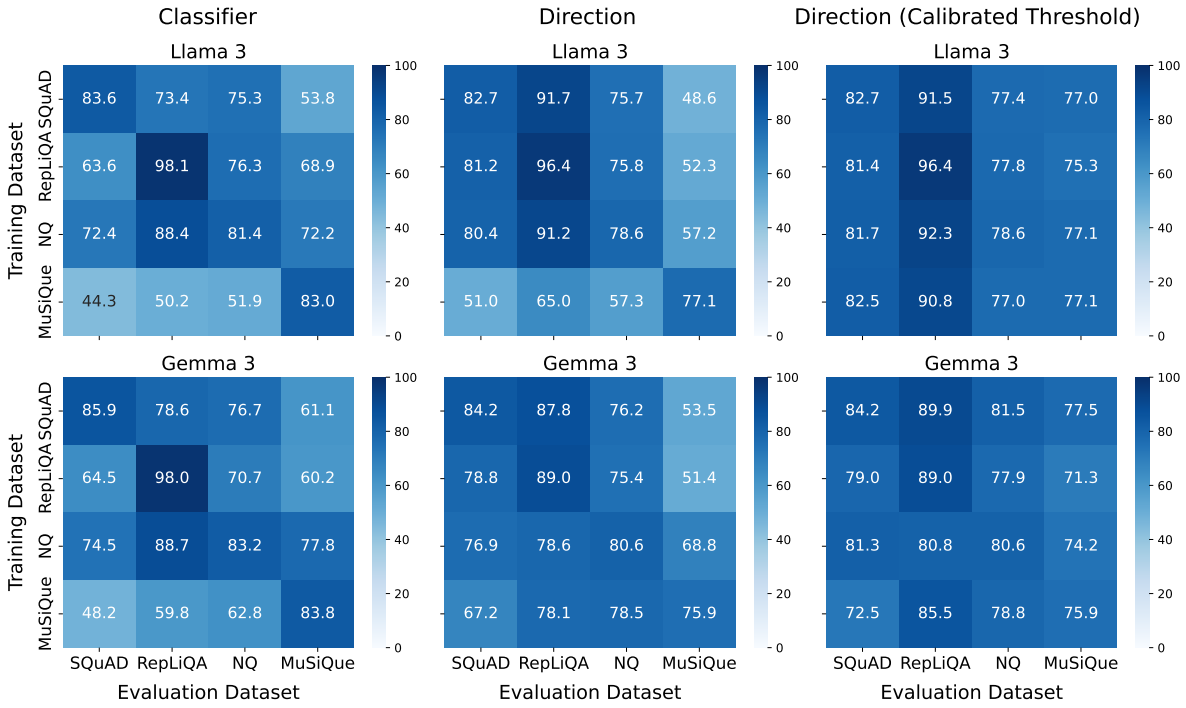

Figure 3: Macro-average F1 scores across datasets using three methods: a trained classifier, a direction-based method with a fixed threshold, and a calibrated threshold variant. Each heatmap shows generalization performance from training on one dataset (rows) to evaluating on another (columns). Results are shown for both Llama-3-8B-Instruct (top) and Gemma-3-12B-IT (bottom).

datasets. Figure 1 shows the recall on unanswerable examples for the Standard Prompt and the Abstain-aware Prompt baselines. Both the classifier and our method outperform the prompt-based baselines. When the training and test splits are from the same dataset, the classifier achieves the highest overall recall, averaging 87% for Llama-3-8B-Instruct and 86.4% for Gemma-3-12B-IT. Our direction-based method is slightly below, with an average recall of 83.6% and 81.9%, respectively. However, when evaluated on unseen datasets, the classifier performance drops by an average of 30.2%, while our method drops by only 7.4%, demonstrating better generalization.

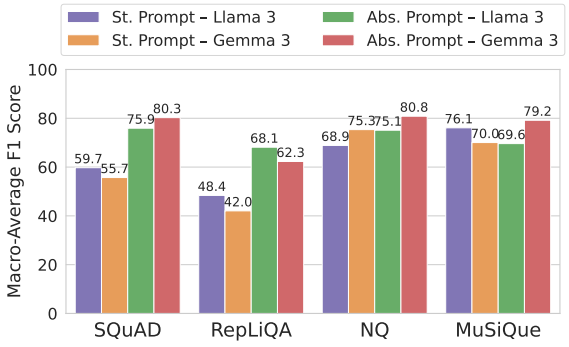

Figure 4: Macro-average F1 scores on answerable and unanswerable questions under Standard and Abstain-aware prompts, evaluated on Llama-3-8B-Instruct and Gemma-3-12B-IT.

Direction-based classification outperforms baselines on unseen datasets Figure 3 (left and middle) presents the macro-average F1 scores for our method and the classifier baseline across all source–evaluation dataset pairs, and Figure 4 reports scores for the prompt-based baselines. We observe the same trends in F1 scores: when the direction is evaluated on the same dataset is was derived from, it achieves 83.7% on average on Llama-3-8B-Instruct and 82.4% on Gemma-3-12B-IT, compared to 86.5% and 87.8%, respectively, for the classifier. However, our method demonstrates stronger generalization than the classifier baseline when evaluated on unseen datasets. Specifically, on SQuAD, RepLiQA, and NQ, it outperforms the classifier by 1.8%–11.9%, averaged per evaluation dataset. The only exception is MuSiQue, where the classifier generalizes better by 8.4–12.2%. We will next show that these results can be improved with a simple threshold calibration, indicating that even in cases where the direction appears not to generalize well, the issue lies in the decision boundary rather than in the quality of the signal itself.

Threshold calibration further improves generalization To understand whether the weaker generalization results reflect that the direction captured a dataset-specific signal, or simply a need for threshold calibration, we visualize the unanswerability scores ϕ_{unans} across datasets (see §C). We observe that the direction consistently induces a separation between answerable and unanswerable examples, however, the optimal decision threshold varies between datasets. This motivates refining the threshold using the validation set of each evaluation dataset, without modifying the learned direction itself, following the procedure described in §3.3.

As shown in Figures 2 and 3, with dataset-specific thresholding, the direction-based method achieves consistent performance across evaluation datasets, regardless of its source. This simple calibration improves generalization results by 2.7–23.7% across evaluation datasets, achieving performance only 2.6% lower on average than that of directions derived from the same datasets. These results suggests that the unanswerability signal captured by the direction is robust and consistently encoded across datasets.

4.3 Steering Effectiveness

To further show that the selected direction captures an unanswerability signal and to observe whether it can influence abstention, we assess its causal impact. To do so, We perform activation space interventions at the chosen layer ℓ^* and token position p^*, for each dataset and model. For a given context–question pair (c, q) formatted with the Abstain-aware Prompt, we modify the hidden activations at layer ℓ^* and position p^* by adding the selected direction, normalized scaled by α:

$$\tilde{\mathbf{h}}^* = \mathbf{h}^* + \alpha\,\hat{\mathbf{v}}^*$$

where $\alpha \in [-2, 2]$ controls the strength and polarity of the intervention. We use GPT-4o mini to determine if the model abstained or attempted to answer the question (see §A.4), and measure the abstention rate on both answerable and unanswerable validation examples under each intervention (see Figure 5). In all cases, increasing α leads to a sharp rise in abstention on both unanswerable and answerable inputs, with mean abstention rates (across all datasets) reaching 96.8% and 95.2%, respectively, at $\alpha = 2.0$. Conversely, when $\alpha = -2.0$, abstention drops to 2.0% for answerable prompts and 19.4% for unanswerable ones. These results

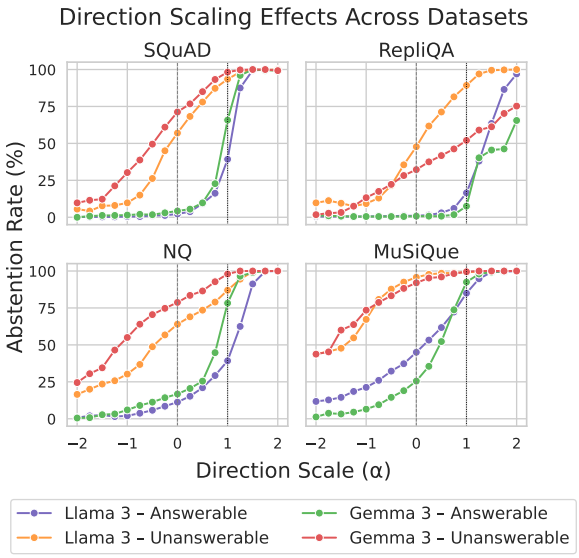

Figure 5: Effect of activation interventions on model abstention rates across steering strengths (α). Results are shown for both answerable and unanswerable validation examples, for each dataset and model.

provide strong evidence that the direction causally influences the model's decision to abstain.

5 Error Analysis

To better understand the limitations of our method, we conducted a manual categorization of 100 misclassified examples: 50 from SQuAD and 50 from NQ, evenly split between answerable and unanswerable instances. Each was assigned to one of five categories:

- *Direction Failure*: the direction score led to an incorrect prediction despite a correct label and well-formed input.
- *Incorrect Label*: the ground-truth annotation appears wrong.
- *Required Title*: (SQuAD only) the document title (not included in our inputs) was necessary to interpret the passage.
- *Grammar "Mistake"*: ungrammatical phrasing or ambiguity made the input difficult to interpret.
- *Answer Not in Context*: the answer exists in the passage but is not clearly in the context of the question.

Table 2 shows the results. We find that 53% of the errors are due to direction failures, and 26%

Category	SQuAD		NQ		Overall
	Ans	Unans	Ans	Unans	%
Direction failure	14	20	7	12	53
Incorrect label	5	3	7	11	26
Required title	3	0	0	0	3
Grammar "mistake"	0	2	2	2	6
Answer not in context	3	0	9	0	12

Table 2: Manual categorization of 100 direction-based classification errors, evenly sampled from SQuAD and NQ (with 25 answerable and 25 unanswerable examples from each).

stem from annotation errors, especially among NQ *unanswerable* examples. Notably, 24% of the misclassified *answerable* examples fall into the "answer not in context" category, most of them in NQ. Overall, this categorization reveals that many of the model's errors arise from ambiguous inputs or limitations in the dataset, rather than clear failures of the method itself. Representative examples from each category are included in §D.

6 Related Work

Prior work has explored methods to improve abstention behavior in models: Lan et al. (2020) improved reasoning with a pretraining loss, leading to improved performance on QA tasks, including unanswerable questions, whereas Zhang et al. (2021) introduced a verification process to detect when questions cannot be answered. Fine-tuned approaches have also been proposed to reduce hallucinations by improving the model's ability to abstain (Zhang et al., 2024; Feng et al., 2024). In contrast, we detect unanswerability by interpreting internal representations of the model, leaving it unchanged.

Several works (Tomani et al., 2024; Kim et al., 2024) evaluated model uncertainty as a signal for whether a question could be answered given the context. We, however, focus directly on unanswerability detection, without estimating uncertainty.

Prompt manipulations were also proposed to detect unanswerability, but showed unstable performance across datasets and models (Slobodkin et al., 2023; Zhou et al., 2023). Slobodkin et al. (2023) further identified an unanswerability-related subspace by training a logistic regression classifier on last-layer hidden representations. Here, we aim to identify a direction in activation space that influences the model's abstention behavior and captures unanswerability consistently across datasets. Another approach used sparse autoencoder features to classify unanswerable inputs (Heindrich et al., 2025). Though effective on the training dataset, the generalization ability of the last two methods proved inconsistent. In contrast, our approach offers a lightweight method for unanswerability classification and demonstrates stronger generalization across datasets.

Extracting linear directions from model activation has been a common technique for analyzing and modifying model behavior (Bolukbasi et al., 2016; Li et al., 2023; Marks and Tegmark, 2024; Hong et al., 2025; Cohen et al., 2025). In this work, we show that similar techniques can be applied to identify a direction associated with unanswerability, and demonstrate how we can use this direction to classify whether a question can be answered from the given context.

7 Conclusion

Our work introduces a method for identifying a direction in the model's activation space that captures unanswerability, using difference-in-means and a selection criterion based on activation steering. We introduce a simple classification method that uses this direction to detect unanswerable questions. We compare our method to existing approaches and find that, while the strongest baseline achieves slightly higher performance when evaluated on its training dataset, our method generalizes more effectively across datasets. We also show that causal interventions along the direction induce abstention behavior of the model. These findings support the view that abstract properties such as unanswerability are linearly encoded in the intermediate representations of language models, and show that this signal can be leveraged for both interpretation and practical use.

Limitations

Our approach assumes that unanswerability is mediated by a linear direction from a fixed layer and token position. While we capture a strong signal, it is possible that unanswerability is represented in more complex patterns, such as across multiple layers or within a circuit, which our method cannot

identify. In addition, we use a simple threshold over the projection onto the direction for classification and do not explore more expressive functions, which could potentially better exploit this signal. Finally, our evaluation is limited to extractive QA tasks. It remains to be seen how well the method extends to other settings, such as open-ended generation.

References

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432, Bangkok, Thailand. Association for Computational Linguistics.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083. Curran Associates, Inc.

Nora Belrose. 2024. Diff-in-means concept editing is worst-case optimal: Explaining a result by sam marks and max tegmark, 2023. *URL https://blog. eleuther. ai/diff-in-means*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Roi Cohen, Omri Fahn, and Gerard de Melo. 2025. Pretrained llms learn multiple types of uncertainty. *Preprint*, arXiv:2505.21218.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

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

Lovis Heindrich, Philip Torr, Fazl Barez, and Veronika Thost. 2025. Do sparse autoencoders generalize? a case study of answerability. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Yihuai Hong, Dian Zhou, Meng Cao, Lei Yu, and Zhijing Jin. 2025. The reasoning-memorization interplay in language models is mediated by a single direction. *Preprint*, arXiv:2503.23084.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Hazel Kim, Adel Bibi, Philip Torr, and Yarin Gal. 2024. Detecting llm hallucination through layer-wise information deficiency: Analysis of unanswerable questions and ambiguous prompts. *Preprint*, arXiv:2412.10246.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry,

and 13 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2024. Zero-resource hallucination prevention for large language models. In *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*, EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024, pages 3586–3602. Association for Computational Linguistics (ACL). Publisher Copyright: © 2024 Association for Computational Linguistics.; 2024 Findings of the Association for Computational Linguistics, EMNLP 2024 ; Conference date: 12-11-2024 Through 16-11-2024.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

João Monteiro, Pierre-André Noël, Étienne Marcotte, Sai Rajeswar, Valentina Zantedeschi, David Vázquez, Nicolas Chapados, Christopher Pal, and Perouz Taslakian. 2024. Repliqa: A question-answering dataset for benchmarking llms on unseen reference content. In *Advances in Neural Information Processing Systems*, volume 37, pages 24242–24276. Curran Associates, Inc.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *Preprint*, arXiv:2310.15154.

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. 2024. Uncertainty-based abstention in llms improves safety and reduces hallucinations. *Preprint*, arXiv:2404.10960.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

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

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14506–14514.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Experimental Setup - Additional Details

This section provides additional details about our experimental setup, including further details on the curated datasets and prompt templates used in our experiments

A.1 Curated Versions of NQ and MuSiQue

We use the curated versions of NQ and MuSiQue introduced by Slobodkin et al. (2023). In NQ, each example consists of a real user question paired with a paragraph from a Wikipedia article. Answerable instances are drawn from questions that include both a long and short answer; the long answer is used as context. Unanswerable instances are constructed by replacing the context with a semantically similar paragraph from the same article that is *not* annotated as the long answer. Paragraphs are ranked using cosine similarity over Sentence-BERT embeddings.

Model	Chat Template
Llama-3-8B-Instruct	<|start_header_id|>user<|end_header_id|>{instruction}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Gemma-3-12B-IT	<start_of_turn>user {instruction}<end_of_turn><start_of_turn>model

Table 3: Chat templates used to format the user instruction during inference.

MuSiQue is a multi-hop QA benchmark in which each instance includes a complex question, a decomposition into sub-questions, and a set of candidate paragraphs. In the curated version, answerable examples are formed by concatenating the gold paragraphs aligned with each sub-question. To generate unanswerable examples, one or more of these gold paragraphs are replaced with the most semantically similar but incorrect paragraphs, identified using the same retrieval method applied in NQ.

A.2 Model Chat Templates

Llama-3-8B-Instruct and Gemma-3-12B-IT are instruction-tuned using system-defined chat templates that wrap the user instruction before response generation.we use these same templates in our experiments, as shown in Table 3. As described in §4.1, we extract hidden activations at the template positions following the user instruction.

A.3 Prompt-based Baseline Prompts

Table 4 shows the prompt used in the standard prompt-based baseline, which contains only the context and question. Table 5 presents the modified version used in the abstention-instruction baseline, which encourages the model to abstain when the question cannot be answered from the passage.

Given the following passage and question, answer the question.

Passage: <passage>

Question: <question>

Answer:

Table 4: Prompt used in the standard prompt-based baseline.

Given the following passage and question, answer the question.

First make sure if it can be answered by the passage.

If it cannot be answered based on the passage, reply "unanswerable".

Passage: <passage>

Question: <question>

Answer:

Table 5: Prompt used in the abstention-instruction baseline.

A.4 Evaluating Prompt-Based Baselines with GPT-4o mini

Table 6 displays the full prompt given to GPT-4o-mini to determine whether a model's response constitutes an abstention. The prompt includes detailed instructions and few-shot examples. To assess the reliability of this evaluation method, we conducted a manual evaluation over 50 model responses: 25 express abstention and 25 attempt to answer. The responses were sampled from model outputs generated for inputs from our datasets and labeled manually. GPT-4o-mini correctly classified all 50 examples.

A.5 Direction and Threshold Selection

Table 7 shows the layer and token position selected for each model–dataset pair, based on the method described in §3.2. For Llama-3-8B-Instruct, the same layer and position (layer 16, position -1) were selected across all datasets. For Gemma-3-12B-IT, the selected layers range from 26 to 27, with positions -1 or -4, depending on the dataset. Since Llama-3-8B-Instruct has 32 layers and Gemma-3-12B-IT has 48, the selected layers lie near the middle of each model.

We compute ROC curves on the validation sets (Figure 6) to select classification thresholds based on the separation between answerable and unanswerable examples. The chosen threshold minimizes the distance to the ideal point and is applied at test time.

B Full Classification Results

We report the full classification results in Tables 8, 9, 10, and 11. For each direction, derived from SQuAD, RepLiQA, NQ, and MuSiQue, we present precision, recall, and F1 scores for both the *answerable* and *unanswerable* classes, across all evaluation datasets and for both Llama-3-8B-Instruct and Gemma-3-12B-IT. Results are reported both with the original threshold and after applying threshold calibration.

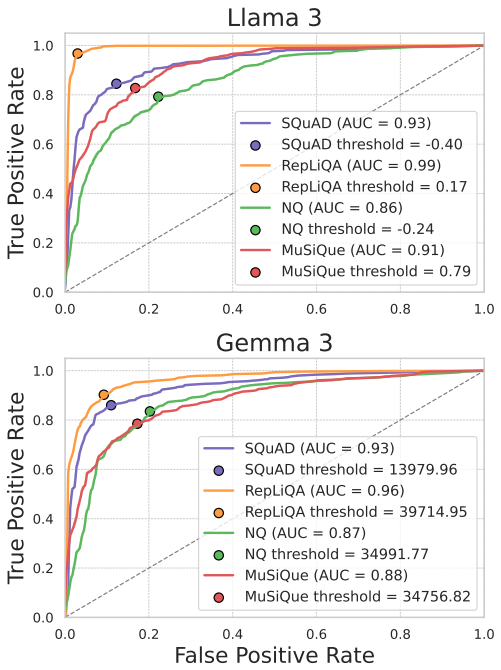

Figure 6: ROC curves across datasets using Llama-3-8B-Instruct and Gemma-3-12B-IT. The x-axis shows the false positive rate (answerable instances classified as unanswerable), and the y-axis shows the true positive rate (unanswerable instances correctly identified). Each curve is annotated with its AUC, and markers indicate the selected classification threshold per dataset.

C Unanswerability Score Distributions

To visualize how well the unanswerability direction separates answerable and unanswerable examples, we plot the unanswerability scores ϕ_{unans}, i.e the results of the projection of hidden activations onto a direction derived from a specific dataset. For each target dataset, we display the distribution of scores for both classes using a fixed direction. Figure 7 show the distributions for Llama-3-8B-Instruct using directions derived from SQuAD and NQ, respectively. In both cases, the direction induces clear separation between the answerable and unanswerable classes. However, the score distributions vary across datasets, suggesting that the optimal decision threshold differs depending on the evaluation set.

You will be shown a model's response to a question. Your task is to decide whether the model attempted to answer the question.

It is considered to abstain if it explicitly says it cannot answer the question, indicates uncertainty, or refuses to provide an answer because the context is insufficient.

Otherwise, it is considered to have answered the question.

Respond with "yes" if the model attempted to answer the question.

Respond with "no" if the model abstained.

You will respond with a single word: "yes" or "no". Do not provide any explanation or extra text.

Examples:

Model response: "1598"

Answer: yes

Model response: "abc"

Answer: yes

Model response: "The answer can be answered by the passage. According to the passage, Fort Coligny was destroyed in 1560."

Answer: yes

Model response: "According to the passage, problems capable of theoretical solutions but consuming unreasonable time in practical application are known as intractable problems"

Answer: yes

Model response: "The passage does not mention BSkyB agreeing with any claim. In fact, it states that BSkyB denied Virgin Media's claim that they had substantially increased the asking price for the channels. Therefore, the answer is ünanswerable."

Answer: no

Model response: "The passage does not answer the question. The passage only mentions that the five counties (Los Angeles, Orange, San Diego, San Bernardino, and Riverside) are among the top 15 most populous counties in the United States, but it does not provide the actual names of the five most populous counties in the United States."

Answer: no

Model response: "The passage does not mention the birth year of Reguier de la Plancha. It only mentions that he died in 1560, but does not provide his birth year."

Answer: no

Table 6: Prompt used to instruct GPT-4o mini to classify model responses as either abstentions or answer attempts.

Model	Dataset	Layer	Position
Llama-3-8B-Instruct	All	16	−1
Gemma-3-12B-IT	SQuAD	27	−1
	RepLiQA	26	−1
	NQ	27	−1
	MuSiQue	27	−4

Table 7: Selected layer and token position from which hidden activations were extracted to compute the unanswerability direction for each model–dataset pair.

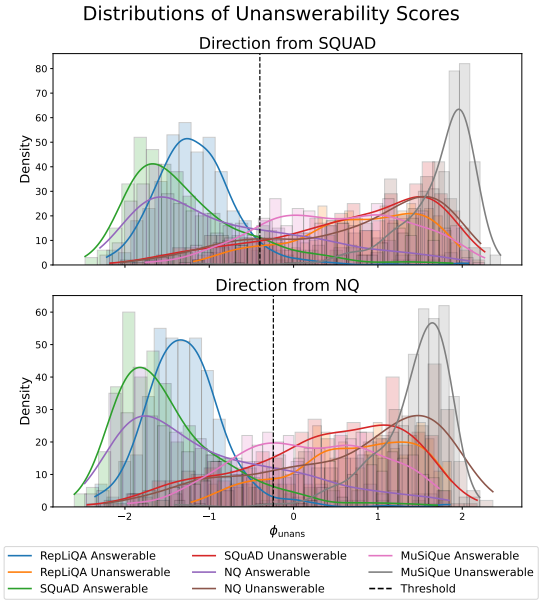

Figure 7: Distribution of unanswerability scores ϕ_{unans} across datasets using the directions derived from SQuAD and NQ in Llama-3-8B-Instruct.

D Failure Case Examples

Table 12 provides one representative example for each of the error categories described in §5. Each row includes the input question, context, predicted label, and a brief explanation of the failure.

E Resources and Packages

In our experiments, we used models and data from the transformers (Wolf et al., 2020) and datasets (Lhoest et al., 2021) packages. AI models (specifically ChatGPT) were used to implement certain helper functions. All the experiments were conducted using a single H100 80GB GPU.

Eval Dataset	Class	Llama 3						Gemma 3					
		Original Threshold			+ Calibration			Original Threshold			+ Calibration		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
SQuAD	Ans	81.2	85.2	83.2	–	–	–	82.5	86.8	84.6	–	–	–
	Unans	84.4	80.3	82.3	–	–	–	86.1	81.6	83.8	–	–	–
RepLiQA	Ans	90.7	92.9	91.8	91.9	90.9	91.4	94.1	80.8	86.9	89.2	90.8	90.0
	Unans	92.7	90.5	91.6	91.0	92.0	91.5	83.2	94.9	88.7	90.6	89.0	89.8
NQ	Ans	81.4	67.1	73.6	78.1	76.1	77.1	86.1	63.4	73.0	80.8	82.5	81.6
	Unans	72.0	84.7	77.9	76.7	78.7	77.7	71.0	89.8	79.3	82.1	80.4	81.3
MuSiQue	Ans	88.7	16.4	27.7	76.8	77.2	77.0	86.8	23.0	36.4	76.7	79.1	77.9
	Unans	53.9	97.9	69.6	77.1	76.7	76.9	55.6	96.5	70.6	78.4	75.9	77.1

Table 8: Full classification results using the direction derived from SQuAD. For each evaluation dataset and class, we report precision (P), recall (R), and F1 score for Llama-3-8B-Instruct and Gemma-3-12B-IT, under the original threshold and after applying threshold calibration.

Eval Dataset	Class	Llama 3						Gemma 3					
		Original Threshold			+ Calibration			Original Threshold			+ Calibration		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
SQuAD	Ans	76.6	90.1	82.8	81.5	81.3	81.4	78.3	79.7	79.0	77.5	81.7	79.6
	Unans	88.0	72.5	79.5	81.3	81.5	81.4	79.3	77.9	78.6	80.7	76.3	78.4
RepLiQA	Ans	96.1	96.7	96.4	–	–	–	90.4	87.4	88.9	–	–	–
	Unans	96.7	96.1	96.4	–	–	–	87.8	90.7	89.2	–	–	–
NQ	Ans	81.3	67.6	73.8	78.8	76.2	77.5	86.4	61.5	71.9	81.2	72.9	76.8
	Unans	72.3	84.4	77.9	77.0	79.5	78.2	70.1	90.3	78.9	75.4	83.1	79.1
MuSiQue	Ans	87.7	21.3	34.3	75.8	74.3	75.1	80.0	21.2	33.5	72.6	68.4	70.4
	Unans	55.2	97.0	70.4	74.8	76.3	75.5	54.6	94.7	69.3	70.1	74.2	72.1

Table 9: Full classification results using the direction derived from RepLiQA. For each evaluation dataset and class, we report precision (P), recall (R), and F1 score for Llama-3-8B-Instruct and Gemma-3-12B-IT, under the original threshold and after applying threshold calibration.

Eval Dataset	Class	Llama 3						Gemma 3					
		Original Threshold			+ Calibration			Original Threshold			+ Calibration		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
SQuAD	Ans	74.7	92.6	82.7	74.2	93.3	82.6	70.5	94.6	80.8	78.5	86.5	82.3
	Unans	90.3	68.7	78.0	91.0	67.5	77.5	91.8	60.5	72.9	85.0	76.3	80.4
RepLiQA	Ans	86.5	97.8	91.8	85.4	97.9	91.2	73.5	90.3	81.0	78.1	85.8	81.8
	Unans	97.5	84.7	90.6	97.5	83.3	89.9	87.4	67.4	76.1	84.2	75.9	79.9
NQ	Ans	79.5	77.3	78.4	–	–	–	82.4	77.9	80.1	–	–	–
	Unans	77.9	80.0	78.9	–	–	–	79.1	83.4	81.2	–	–	–
MuSiQue	Ans	87.6	28.2	42.7	87.5	29.3	43.9	80.6	52.0	63.2	74.2	74.0	74.1
	Unans	57.2	96.0	71.7	57.5	95.8	71.9	64.6	87.5	74.3	74.1	74.3	74.2

Table 10: Full classification results using the direction derived from NQ. For each evaluation dataset and class, we report precision (P), recall (R), and F1 score for Llama-3-8B-Instruct and Gemma-3-12B-IT, under the original threshold and after applying threshold calibration.

Eval Dataset	Class	Llama 3						Gemma 3					
		Original Threshold			+ Calibration			Original Threshold			+ Calibration		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
SQuAD	Ans	55.0	99.6	70.9	79.9	87.0	83.3	62.7	97.3	76.3	69.3	81.4	74.9
	Unans	97.9	18.5	31.1	85.7	78.1	81.7	94.0	42.1	58.2	77.5	64.0	70.1
RepLiQA	Ans	61.3	99.9	76.0	91.0	90.7	90.8	70.8	98.4	82.4	84.7	86.8	85.7
	Unans	99.7	37.0	54.0	90.7	91.0	90.9	97.4	59.5	73.9	86.5	84.3	85.4
NQ	Ans	57.3	96.7	72.0	76.1	78.8	77.4	76.4	82.7	79.4	79.4	78.0	78.7
	Unans	89.5	28.0	42.7	78.0	75.2	76.6	81.1	74.4	77.6	78.4	79.7	79.0
MuSiQue	Ans	77.5	76.5	77.0	–	–	–	74.0	80.0	76.9	–	–	–
	Unans	76.8	77.8	77.3	–	–	–	78.2	71.9	74.9	–	–	–

Table 11: Full classification results using the direction derived from MuSiQue. For each evaluation dataset and class, we report precision (P), recall (R), and F1 score for Llama-3-8B-Instruct and Gemma-3-12B-IT, under the original threshold and after applying threshold calibration

Category	Dataset	Label	Context	Question	Comment
Direction Failure	SQuAD	Unanswerable	The topic of language for writers from Dalmatia and Dubrovnik... These facts undermine the Croatian language proponents' argument that modern-day Croatian is based on a language called Old Croatian.	Prior to the 19th century where did Croatians and Serbians live?	The direction predicts "answerable," but the context passage does not answer the question.
Incorrect Label	NQ	Unanswerable	The ileum is the third and final part of the small intestine... It ends at the ileocecal junction...	Where is the ileum located in the body?	Based on the passage, the ileum is located in the small intestine, specifically as the third and final part of it. Therefore, the label is incorrect.
Required Title	SQuAD	Answerable	During the latter half of the 20th century, a more diverse range of industry also came to the city, including aircraft and car manufacture...	Southampton's range of industries includes the manufacture of cars and what other transport?	Without the passage title ("Southampton"), it is unclear that "the city" refers to the correct subject.
Grammar "Mistake"	SQuAD	Unanswerable	...MCA agent Lew Wasserman made a deal with Universal for his client James Stewart...	Who was a MAC agent?	The question refers to "MAC" while the passage only mentions "MCA" so the correct label is unanswerable. However, the model may have treated this as a minor typo, leading the direction to misclassify it as answerable.
Answer Not in Context	NQ	Answerable	The Speaker, Majority Leader, Minority Leader, Majority Whip and Minority Whip all receive special office suites...	Top 5 leadership positions in the House of Representatives?	The roles are listed, but the passage does not clearly frame them as leadership positions relevant to the question.

Table 12: Representative misclassified examples from each failure category. Each includes the dataset, gold label, and a brief explanation of the error.