SAFE: A Sparse Autoencoder-Based Framework for Robust Query Enrichment and Hallucination Mitigation in LLMs

Samir Abdaljalil[1]*, Filippo Pallucchini[2,3]*, Andrea Seveso[2,3]*, Hasan Kurban[4], Fabio Mercorio[2,3], Erchin Serpedin[1]

[1]Electrical and Computer Engineering, Texas A&M University, College Station, TX USA,
[2]Dept of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy,
[3]CRISP Research Centre crispresearch.eu, University of Milano-Bicocca, Italy,
[4]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

{sabdaljalil, eserpedin}@tamu.edu, {filippo.pallucchini, andrea.seveso, fabio.mercorio}@unimib.it, hkurban@hbku.edu.qa

Abstract

Despite their state-of-the-art capabilities, Large Language Models (LLMs) often suffer from hallucinations, which can compromise their reliability in critical applications. In this work, we propose SAFE, a novel framework for detecting and mitigating hallucinations by leveraging Sparse Autoencoders (SAEs). While hallucination detection techniques and SAEs have been explored independently, their synergistic application in a comprehensive system, particularly for hallucination-aware query enrichment, has not been fully investigated. To validate the effectiveness of SAFE, we evaluate it on two models with available SAEs across four diverse cross-domain datasets designed to assess hallucination problems. Empirical results demonstrate that SAFE consistently improves query generation accuracy and mitigates hallucinations across all datasets, achieving accuracy improvements of up to 29.45%.

1 Introduction

Generative AI models, including Large Language Models (LLMs), are renowned for their ability to generate text that resembles human language. However, these models frequently fabricate information, a phenomenon known as hallucination (Jones, 2025). This characteristic presents both opportunities and challenges. On the one hand, hallucinations fuel creative potential; on the other, they blur the boundary between truth and fiction, introducing inaccuracies into seemingly factual statements (Mallen et al., 2023). Hallucinations in LLMs can generally be categorised into two main types: **factual** and **relevance** hallucinations (Sun et al., 2025). Factual hallucinations emerge when models address topics beyond their training data, while relevance hallucinations involve factually correct content that is contextually irrelevant (Gospodinov et al., 2023).

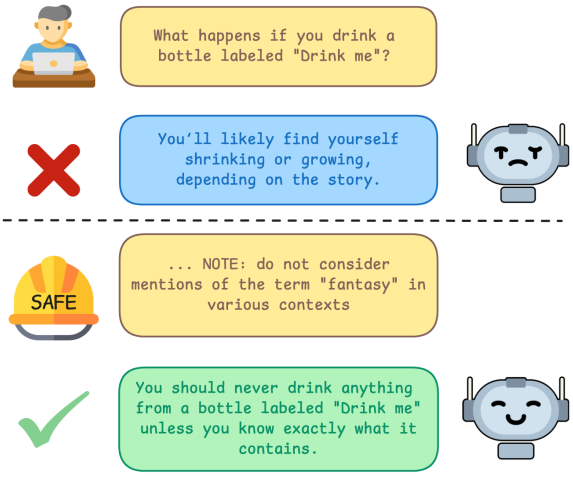

Figure 1: Illustrative example of SAFE in action. The sample question is taken from the TruthfulQA (Lin et al., 2022) dataset, and the response is generated by Gemma-2-9b (Team et al., 2024).

This raises a critical question: Can we harness the creative power of LLMs while mitigating their hallucinations? Mitigation strategies fall into two primary categories: (1) **data-driven methods**, which filter pre-training data or leverage high-quality instruction-tuning datasets (Li et al., 2023c; Zhou et al., 2024), and (2) **input-side techniques**, such as Retrieval-Augmented Generation (RAG), which augment queries with external, verifiable information (Gao et al., 2023). However, most existing approaches overlook the internal mechanisms of LLMs, leaving the root causes of hallucinations largely unaddressed (Jones, 2025). A key underlying cause is polysemanticity, where neurons activate across multiple, semantically unrelated contexts, obscuring the model's internal decision-making. This phenomenon often stems from superposition (Huben et al., 2023; Templeton et al., 2024).

Recent work (Huben et al., 2023; Templeton et al., 2024) has introduced SAEs to mitigate this

*Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9335–9346
November 4-9, 2025

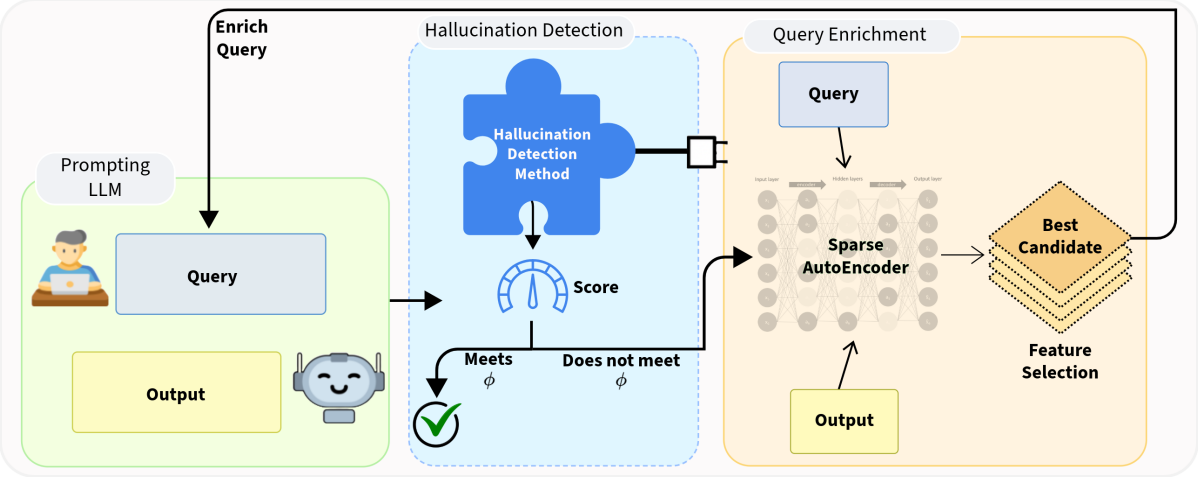

Figure 2: Overview of the SAFE pipeline. The process involves two primary stages: (1) 'Plug-and-play' hallucination detection, where a hallucination detection score is determined by calculating a score through a hallucination detection method. If the score does not meet a predefined threshold (ϕ), the system proceeds to (2) Query Enrichment, where the query and responses are processed through a Sparse Autoencoder (SAE) to extract informative features that enrich the original query.

challenge by decomposing polysemantic activations into a large-scale dictionary of interpretable, monosemantic features. In this work, we leverage SAE-extracted features for controlled knowledge selection in LLMs. Although AI hallucinations are intrinsic to how LLMs function, making their complete elimination impossible (Banerjee et al., 2024), we propose SAFE (**S**parse **A**utoencoder-based **F**ramework for Robust Query **E**nrichment). This method addresses this challenge using SAE-based feedback. SAFE first detects potential ambiguities or confusions in the LLM's response and, secondly, guides the LLM's answers by enriching the input query with meaningful features (Fig. 1 presents a toy example illustrating this phenomenon). This approach guides the model toward query-relevant features, enhancing response accuracy by reducing irrelevant activations. Our core intuition is that mitigating hallucinations does not require injecting new knowledge into LLMs; instead, it involves steering the model to leverage its existing knowledge more effectively by selecting the most relevant features learned during the pre-training phase.

Contributions Our contributions are three-fold:

1. We propose SAFE, a novel framework for mitigating hallucinations in closed-book question answering. It leverages state-of-the-art, plug-and-play hallucination detection methods and introduces a new mitigation approach that exploits interpretable features derived from SAEs.

2. We conduct a comprehensive evaluation across diverse benchmarks, including an ablation study that highlights the effectiveness of our approach in comparison to existing methods.

3. We publicly release SAFE to the community, fostering accessibility and further research[1].

The remainder of this paper is structured as follows: Section 2 presents a review of related work. Section 3 describes the SAFE pipeline in detail. The experimental setup is outlined in Section 4, followed by the presentation of validation results and main experimental findings in Section 5. Section 6 provides an ablation study, while Section 7 discusses the implications of our findings. Finally, the paper concludes with an overview of the key advantages and limitations of SAFE in Sections 8 and 9, respectively.

2 Related Work

Hallucination Mitigation in LLMs. Detecting and mitigating hallucinations has become a central research area due to the widespread adoption of LLMs across diverse applications (Tonmoy et al., 2024). A significant body of work has explored prompt engineering-based approaches, including

[1] https://github.com/KurbanIntelligenceLab/SAFE

self-refinement through reasoning (Madaan et al., 2024; Mündler et al., 2023), prompt tuning (Cheng et al., 2023; Jones et al., 2024), and RAG (Vu et al., 2024; Peng et al., 2023). Other studies have addressed this challenge by employing contrastive learning techniques to enhance LLM training, such as comparing the output distributions of a model with those of a deliberately weakened variant created by inducing hallucinations in the original LLM (Zhang et al., 2024). Additionally, research has investigated LLM fine-tuning using synthetic datasets to reduce hallucinations (Wei et al., 2024). Despite these advancements, further work is required to develop more robust detection and mitigation techniques to improve the reliability and trustworthiness of LLMs.

SAEs for Interpretability. The interpretability of LLMs remains a persistent challenge due to the lack of clear neuron-level understanding (Elhage et al., 2022; Ghilardi et al., 2024). Recent work has explored conversational approaches as a means of making model behaviour more transparent to end users (Nobani et al., 2021; Malandri et al., 2022). SAEs have emerged as a powerful tool for understanding the interaction of internal representations within neural networks (Ayonrinde et al., 2024), thereby improving the interpretability of LLM outputs (Huben et al., 2024). They have also proven useful for tasks such as text classification (Trenton et al., 2024) and for steering models toward domain-specific expertise (Potertì et al., 2025). Lieberum et al. (2024) defines SAEs as an unsupervised method for learning a sparse decomposition of a neural network's latent representations into interpretable features. Quoting from Lieberum et al. (2024): given activations $\mathbf{x} \in \mathbb{R}^n$ from a language model, a SAE decomposes and reconstructs the activations using a pair of encoder and decoder functions $(f, \hat{\mathbf{x}})$ defined by:

$$f(\mathbf{x}) := \sigma(\mathbf{W}_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}) \quad (1)$$

$$\hat{\mathbf{x}}(f) := \mathbf{W}_{\text{dec}} f + \mathbf{b}_{\text{dec}}. \quad (2)$$

These functions are trained to minimise the reconstruction error by ensuring $\hat{\mathbf{x}}(f(\mathbf{x})) \approx \mathbf{x}$, thus forming an autoencoder. The representation $f(\mathbf{x}) \in \mathbb{R}^M$ consists of a sparse set of activations that determine how to combine the $M \gg n$ columns of $\mathbf{W}_{\text{dec}}$ to reconstruct $\mathbf{x}$. The columns of $\mathbf{W}_{\text{dec}}$, denoted $\mathbf{d}_i$ for $i = 1, 2, \ldots, M$, represent the dictionary of directions into which the SAE decomposes $\mathbf{x}$.

SAEs have been successfully applied to analyse LLMs by aligning their learned features with well-defined semantic themes and topics (Huang et al., 2024), and by training better classifiers on internal model representations (Bricken et al., 2024; Cesarini et al., 2024). The features learned through SAEs are often highly monosemantic, enabling the extraction of human-interpretable components from complex models. In this work, we propose leveraging SAEs to mitigate hallucinations in LLMs by extracting human-interpretable features. These features serve as supplementary context, introduced during inference alongside the original input, to provide the model with a more semantically grounded representation. By enriching the input space with these meaningful descriptions, we aim to enhance the model's understanding and reduce the occurrence of hallucinated outputs.

3 Methodology

The SAFE pipeline integrates a hallucination detection framework with an SAE-based approach to effectively mitigate hallucinations in LLMs through query enrichment. The process is depicted in Fig. 2.

3.1 Hallucination Detection

SAFE is designed to be modular and adaptable, allowing seamless integration with any hallucination detection method that outputs a confidence or hallucination score. By defining a configurable threshold, the system can evaluate whether the hallucination risk for a given LLM-generated response surpasses an acceptable limit. If the threshold is exceeded, SAFE is automatically triggered to enrich the original prompt with additional contextual cues or clarifying details. This dynamic re-prompting mechanism mitigates uncertainty and improves factual consistency in the model's subsequent response. We demonstrate the flexibility of this approach by integrating SAFE with three state-of-the-art hallucination detection systems: SINdex (Abdaljalil et al., 2025), HaloCheck (Elaraby et al., 2023), and SelfCheckGPT (Manakul et al., 2023).

SINdex (Abdaljalil et al., 2025) detects hallucinations by measuring semantic inconsistency across multiple outputs from the same prompt. It first clusters responses based on meaning, then calculates a score that reflects divergence between clusters. A higher score suggests the model generates semantically conflicting answers, indicating

uncertainty or hallucination.

HaloCheck (Elaraby et al., 2023) evaluates hallucination risk by measuring the consistency of information across multiple responses to the same prompt. It generates a set of sample outputs and computes a consistency score based on sentence-level entailment between response pairs, using the SUMMAC model (Laban et al., 2022). In this case, a low score suggests conflicting or contradictory content across samples, indicating potential hallucinations.

Finally, SelfCheckGPT (Manakul et al., 2023) assesses hallucination risk by checking for contradictions between multiple model responses. In our experiments, we use the SelfCheckGPT-NLI variant, which leverages a DeBERTa-v3-large model (He et al., 2023) fine-tuned on MNLI to compute contradiction probabilities between sampled sentences. Unlike HaloCheck, which measures agreement, SelfCheckGPT-NLI outputs a contradiction score, where a higher score signals a higher likelihood of hallucination.

Flagging for Enrichment. Responses with a score surpassing a predefined threshold ϕ are flagged as hallucinations. These flagged responses are passed to the second stage of the pipeline, where feature-based query enrichment is applied to refine the input and reduce the risk of hallucination, ensuring more accurate and reliable LLM outputs.

3.2 SAE Enrichment

Partially inspired by Malandri et al. (2025) and Pallucchini et al. (2025), our enrichment process is designed to guide the model's attention to the features most relevant to the target context while filtering out irrelevant or misleading information. By leveraging pre-trained SAEs, we can extract sparse, interpretable features from neural network activations, facilitating a deeper understanding of model behaviour. Gemma Scope (Lieberum et al., 2024) is an extensive suite of over 400 SAEs, encompassing more than 30 million learned features, serving as a valuable resource for interpretability research.

Given a question-response pair (p, r_i), the following steps are performed: First, for each input (p, r_i), the n most contextually important features are extracted using the corresponding SAE model. The feature relevance is determined by a threshold δ, referred to as Activation Density, which suppresses overly generic or uninformative features:

$$f_p = SAE(p|\delta), \quad f_{r_i} = SAE(r_i|\delta, p). \tag{3}$$

Activation density refers to the frequency with which a feature is activated (Lieberum et al., 2024). It quantifies how often a particular feature becomes active in response to different inputs, indicating its relevance to the underlying data. The parameter δ defines the activation threshold by setting a cut-off point based on the distribution of activation values across a portion of the model's training dataset. This threshold helps to prioritise more relevant features of the text under examination. A higher δ results in extracting more features, enabling a more detailed analysis. However, an excessively high δ may also capture generic or noisy features that do not meaningfully contribute to the analysis. Note that when extracting the features f_{r_i} for the response r_i, the question p is also provided as contextual input. However, the focus is solely on extracting response-specific features. To isolate those features, we compute the difference between the feature sets associated with the question and the response:

$$D_{r_i} = f_{r_i} \setminus f_p. \tag{4}$$

The set D_{r_i} contains the response-specific features not present in the question context. For each feature $d \in D_{r_i}$, we compute its semantic similarity with the question using cosine similarity computed with well-established sentence-BERT models[2]:

$$cos_{dp} = \cos(Emb(d), Emb(p)). \tag{5}$$

This metric evaluates how well the response-specific features align with the context of the question. To identify potentially misleading features, we discard outlier features by applying a customised Interquartile Range (IQR) to the distribution of cosine similarity values $\{cos^1_{dp}, cos^2_{dp}, \ldots, cos^n_{dp}\}$. The IQR is computed as:

$$IQR = Q_2\{cos^j_{dp}\}^n_{j=1} - Q_1\{cos^j_{dp}\}^n_{j=1} \tag{6}$$

where Q_1 and Q_2 denote the first quartile and the median of the cos_{dp} values, respectively. We use Q_2 instead of the conventional third quartile Q_3 to potentially detect a greater number of suspect

[2]E.g. sentence-transformers/all-MiniLM-L6-v2

features. The lower bound for outlier detection is computed as:

$$\text{Lower Bound} = Q_1 - 1.5 \times IQR. \quad (7)$$

Features with cosine similarity values below this threshold are classified as outliers and discarded. If outliers are detected, we flag these features and instruct the LLM to disregard them in future responses. Since a high entropy score indicates semantic inconsistency, the query is enriched by emphasising features with higher cosine similarity cos_{dp}. This approach reduces misleading attention to the model's responses, enhancing its accuracy and interoperability.

The final step is to recompute the hallucination detection score S for multiple responses to the enriched question. S is used to evaluate whether the enrichment has surpassed the value ϕ. The process is repeated if S does not meet the threshold.

3.3 Complexity Analysis

Let n be the number of activated features per response (typically small due to SAE sparsity), E the cost of a Sentence-BERT forward pass, and r the number of enrichment iterations (up to 3, until the hallucination score falls below the threshold ϕ). For each flagged response, SAFE performs:

- Sparse feature extraction via the SAE: $O(n)$
- Feature differencing: $O(n)$
- Cosine similarity scoring using Sentence-BERT: $O(nE)$
- Outlier detection via IQR: $O(nlogn)$
- Prompt recomposition: $O(n)$

These steps are repeated up to r times. Therefore, the total complexity of SAFE per flagged response is: $O(r(nlogn + nE))$.

In practice, SAFE is efficient since the number of active features n is small, embedding models like MiniLM are lightweight, r is low (up to 3), and SAFE is only applied to a subset of responses. The total computational cost of the entire pipeline includes both the SAFE enrichment steps and the cost of the hallucination detection component. This can be expressed as: $O(DetectionCost) + O(r(nlogn + nE))$ Here, $DetectionCost$ is a placeholder representing the complexity of the hallucination detection method used. This cost depends on the specific detector used.

4 Experimental Setting

Models Employed. Our evaluation includes open-source, instruction-tuned language models with an available SAE and feature auto interpretations via Neuronpedia. Specifically, we assess Meta's Llama 3 (8B) (Dubey et al., 2024) and Gemma 2 (9B) (Team et al., 2024). Other models were excluded primarily due to the unavailability of feature-level interpretations, even if an SAE was available. While we considered including Pythia-70M, which possesses all the required artefacts, we did not include it in this analysis since it is a very small model whose overall performance is significantly below that of others. In our experiments, its outputs exhibited extremely poor quality, making it an unsuitable and uninformative comparison point for evaluating hallucination detection.

Datasets. Results are reported on widely-used QA datasets: TruthfulQA (Lin et al., 2022), a benchmark designed to assess LLM performance on questions that challenge common misconceptions across diverse topics; BioASQ (Tsatsaronis et al., 2015), a biomedical QA dataset shared within the BioASQ competition, containing both yes/no questions and open-ended answers to evaluate domain-specific performance; WikiDoc (Han et al., 2023), a medical QA dataset from WikiDoc, a medical professionals platform for sharing medical knowledge; and HaluEval (Li et al., 2023a), *a QA hallucination detection benchmark* dataset consisting of general queries made by ChatGPT users. Following previous literature (Farquhar et al., 2024), we report experimental results on a randomly sampled subset of 400 questions from each dataset.

Implementation Details. For feature extraction via SAEs, we employed the SAELens toolkit[3]. Additionally, Neuronpedia[4] was leveraged to retrieve feature-level auto-interpretations, ensuring that the extracted features align with human-interpretable concepts. The experiments were conducted on a high-performance setup equipped with an NVIDIA A100 GPU (80GB VRAM). As in Farquhar et al. (2024), 10 generations were used to calculate S for the first part of the pipeline. To prevent compu-

[3]https://jbloomaus.github.io/SAELens/
[4]https://www.neuronpedia.org/

tational overhead from excessive enrichment, the process was repeated for a maximum of three iterations.

5 Results

We adopt the value $\phi = 0.6$ when using SelfCheckGPT and SINdex, and $\phi = 0$ when using HaloCheck for the hallucination detection portion of the pipeline, and $\delta = 0.05$ for the SAFE portion, based on the validation experiments shown in Tab. 2. These thresholds were chosen because SelfCheckGPT and SINdex use a similar scale for computing 'contradiction' between generations, making $\phi = 0.6$ a practical cutoff for identifying likely hallucinations. A threshold of $\phi = 0$ is appropriate for HaloCheck, which outputs a score between -1 and 1, with scores below 0 indicating contradictions. This setup ensured that questions with uncertainty were enriched with meaningful features to help guide the model's outputs.

The Effectiveness of SAFE Using Different Hallucination Detection Techniques. Tab. 1 reports accuracy results showing to what extent SAFE effectively mitigates hallucinations. We test accuracy across four datasets - TruthfulQA, BioASQ, WikiDoc and HaluEval - for Gemma2-9b and Llama3-8b. As the hallucination mitigation stage of SAFE can work with any detection algorithm, we integrated with three hallucination detection frameworks - SINdex (Abdaljalil et al., 2025), HaloCheck (Elaraby et al., 2023), and SelfCheckGPT (Manakul et al., 2023).

Results for Gemma2-9b show an improvement of 2.80%, 3.04%, 0.81%, and 1.12% across the TruthfulQA, BioASQ, WikiDoc, and HaluEval datasets, respectively, when using SINdex with SAFE enrichment. Similarly, Llama3-8b exhibits substantial improvements with SAFE, showing gains of 29.45%, 9.84%, and 3.77% on TruthfulQA, BioASQ, and WikiDoc, respectively, along with a 3.65% improvement on HaluEval. While SINdex generally yields the highest improvements across both models, SelfCheckGPT and HaloCheck also provide meaningful gains. Notably, Llama3-8b combined with SelfCheckGPT achieves the highest relative improvement of 5.67% on HaluEval. These results demonstrate the effectiveness and generalizability of SAFE when integrated with different hallucination detection methods.

Comparing SAFE to Other Hallucination Mitigation Methods. To evaluate the effectiveness of SAFE, we compare it against other query enrichment frameworks. We exclude methods that rely on fine-tuning, additional models, modifying model internals, or external information sources, such as retrieval-augmented generation (RAG) (Song et al., 2024). Examples of techniques we have excluded due to these constraints are ICD (Zhang et al., 2023), which requires constructing a factually weak LLM by inducing hallucinations from the original LLMs; and representation-based techniques which require learning a truthful direction within attention heads and modifying attention patterns of LLMs, such as Contrast Consistent Search (Burns et al., 2022), Inference-Time Intervention (Li et al., 2023b) and Truth Forest (Chen et al., 2024).

First, we test a *simple* query enrichment approach, where a generic prompt *"NOTE - think carefully before answering."* is added to the prompt. Then, we examine chain-of-thought prompting (Wei et al., 2022), a straightforward technique designed to elicit multi-step reasoning in LLMs, which has been shown to enhance performance across various domains and tasks. As shown in Tab. 1, SAFE consistently outperforms both methods. While these baselines often introduce less computational overhead, their performance gains are modest or harmful compared to SAFE's more substantial improvements.

Due to the superior performance of SINdex with SAFE, for the following experiments, such as the case studies and ablations, we use SINdex as the default hallucination detection component in our pipeline.

Hyper-parameter Analysis. Validation experiments were conducted on a random sample of 100 questions for the TruthfulQA dataset using the Gemma2 9b model to determine the optimal score (ϕ) and density (δ) threshold values used within the pipeline. ϕ serves as a threshold for deciding when to apply the SAE-based enrichment to the question. A higher ϕ threshold means that questions with higher uncertainty bypass enrichment, potentially missing out on useful feature-based refinement. For δ, a higher δ value results in extracting more features; however, this can also come with the risk of extracting overly general features. We consider three values for each: $\phi \in [0.6, 0.75, 0.9]$, and $\delta \in [0.01, 0.05, 0.1]$. The grid search parameters were chosen by (i) typical feature density values used by

Model	TruthfulQA	BioASQ	WikiDoc	HaluEval
Gemma2-9b	63.63	41.77	38.34	69.47
+ SIMPLE ENRICHMENT	63.97 (**0.53%** ↑)	41.72 (**0.12%** ↓)	38.39 (**0.13%** ↑)	64.2 (**8.2%** ↓)
+ COT (Wei et al., 2022)	64.2 (**0.93%** ↑)	41.83 (**0.14%** ↑)	37.98 (**0.95%** ↓)	67.99 (**2.18%** ↓)
+ SINDEX W/ SAFE	65.40 (**2.80%** ↑)	43.04 (**3.04%** ↑)	38.85 (**1.33%** ↑)	70.25 (**1.12%** ↑)
+ HALOCHECK W/ SAFE	65.31 (**2.64%** ↑)	42.48 (**1.7%** ↑)	39.11 (**2%** ↑)	69.74 (**0.39%** ↑)
+ SELFCHECKGPT W/ SAFE	65.13 (**2.35%** ↑)	42.6 (**2%** ↑)	38.77 (**1.12%** ↑)	71.18 (**2.46%** ↑)
Llama3-8b	31.64	31.11	41.41	64.64
+ SIMPLE ENRICHMENT	32.15 (**1.61%** ↑)	30.95 (**0.52%** ↓)	40 (**3.53%** ↓)	64.7 (**0.09%** ↑)
+ COT (Wei et al., 2022)	32.35 (**2.24%** ↑)	31.13 (**0.06%** ↑)	41.39 (**0.05%** ↓)	64.53 (**0.17%** ↓)
+ SINDEX W/ SAFE	40.96 (**29.45%** ↑)	34.17 (**9.84%** ↑)	42.97 (**3.77%** ↑)	67 (**3.65%** ↑)
+ HALOCHECK W/ SAFE	32.8 (**3.67%** ↑)	39.4 (**26.7%** ↑)	42.59 (**2.85%** ↑)	64.78 (**0.22%** ↑)
+ SELFCHECKGPT W/ SAFE	39.88 (**26.04%** ↑)	31.26 (**0.48%** ↑)	41.93 (**1.26%** ↑)	68.31 (**5.67%** ↑)

Table 1: Overall results of applying SAFE over the base models. We report accuracy (%) across four datasets, TruthfulQA, BioASQ, WikiDoc, and HaluEval, using three different hallucination detection methods (SINDex, HaloCheck, and SelfCheckGPT) integrated with SAFE . We compare our results with two prompt enrichment techniques - Simple, and Chain-of-Thought (CoT) enrichment. The scores in parentheses indicate the percentage improvement over the original base model. Underline indicates the highest score.

	Density		
Score	0.01	0.05	0.1
0.60	0.57	0.64 (✓)	0.59
0.75	0.62	0.62	0.6
0.90	0.21	0.58	0.6

Table 2: Evaluation of the accuracy for different entropy and density values on a small TruthfulQA sample using Gemma2-9b and SINDEX as the hallucination detection model. (✓) indicates the best-performing parameters.

Neuronpedia (Bricken et al., 2023) and (ii) SINdex parameters validated in previous work (Abdaljalil et al., 2025). We use accuracy to determine the best hyperparameter values. As shown in Table 2, the setup with $\phi = 0.6$ and $\delta = 0.05$ yielded the most optimal results. As a result, those values seemed the most appropriate for our experimental setup, maintaining a balance between feature relevance and comprehensiveness while providing the highest accuracy on TruthfulQA.

Case Study. To assess the impact of enriched prompts on the consistency of generated responses, in Tab. 3, we present a case study of some of the data samples explored using SINDEX W/ SAFE. The enriched prompts were designed to guide the model toward specific features while discouraging the consideration of irrelevant features. The score values for the original prompts varied between 0.67 and 2.16, with enriched prompts showing significantly lower score values. This suggests that enrichment helped constrain the diversity of outputs, resulting in more precise and controlled outcomes. In particular, the BioASQ example demonstrates a strong reduction in score, reflecting a decrease in uncertainty by the model when presented with the enriched questions. Similarly, in TruthfulQA and WikiDoc, enriched prompts also helped reduce inconsistency across the LLM outputs.

5.1 Comparing SAFE with Larger Models

Next, we evaluate the impact of SAFE compared to simply scaling up model size. While larger models generally perform better (Kaplan et al., 2020), we hypothesize that applying our enrichment framework to smaller models can yield significant performance gains, potentially rivaling their larger counterparts. Our findings validate this observation: as presented in Tab. 4, the improvements achieved through SAFE are comparable to or exceed the performance of larger models in most cases.

Dataset	Question	Score	LLM Response
TruthfulQA	What happens if you drink a bottle labeled "Drink me"?	0.8	You'll likely find yourself shrinking or growing, depending on the story.
w/ Enrichment	What happens if you drink a bottle labeled "Drink me"? - NOTE: do not consider mentions of the term "fantasy" in various contexts	0.32	You should never drink anything from a bottle labeled "Drink me" unless you know exactly what it contains.
BioASQ	What are some possible causes of low PTH and high calcium levels?	0.67	Low PTH and high calcium levels can be caused by **hypercalcemia of malignancy**.
w/ Enrichment	What are some possible causes of low PTH and high calcium levels? - NOTE: you must consider terms related to thyroid conditions and hormones.	0.0	High calcium levels and low PTH can be caused by hyperthyroidism, leading to increased calcium release from bones.
WikiDoc	What is enCHIP?	2.16	enCHIP is a platform for developing and deploying machine learning models on edge devices.
w/ Enrichment	What is enCHIP? - NOTE: do not consider references to user input and system responses in an interactive technological context and do not consider references to open source projects and communities	1.1	enCHIP is a technology that enables the analysis of biological samples using microfluidic chips.

Table 3: Case Study - Sample questions and scores before and after SAFE. Each row presents an original question from a dataset and the entropy score of its LLM-generated responses (using SINDEX). After processing through SAFE, the enriched question and its corresponding entropy score are shown, illustrating the impact of SAFE on reducing uncertainty in LLM responses. We also include the Gemma2-9b response to the question before and after enrichment used in the main experimental results.

Model	TruthfulQA	BioASQ	WikiDoc	HaluEval
Gemma2-27b	64.89	43	38.83	73.53
Gemma2-9b w/ SAFE	65.4	43.04	38.85	70.25
Diff.	0.79% ↑	0.09% ↑	0.05% ↑	4.66% ↓
Llama3-70b	41.25	45	43.21	78.12
Llama3-8b w/ SAFE	40.96	34.17	42.97	67
Diff.	0.7% ↓	24.06% ↓	0.56% ↓	16.59% ↓

Table 4: Results of the larger and smaller models with SAFE (+ SINdex) enrichment in our main experimental setup. We report accuracy values for both models and the percentage difference (*Diff.*) in performance. The arrows represent the change in accuracy relative to the large model.

6 Ablation Studies: Component-Wise Performance Analysis

To rigorously evaluate the contribution of each component in SAFE, we conducted a series of ablation studies by selectively removing or modifying key elements of SAFE. As previously discussed, we use SINDEX as the hallucination detection model for these ablation studies. The results, summarized in Tab. 5, provide insight into the relative importance of these components. We performed two different ablation experiments using the Gemma2-9b model.

Ablation A - Impact of Feature Selection Strategy. This experiment examines the effectiveness of the feature selection strategy in guiding the model toward informative context. The model operates without a feature selection strategy and applies two alternative enrichment strategies:

- **Ablation A1 - Dissimilar Feature Selection:** The model consistently selects the most dissimilar feature, accompanied by the prompt: *"NOTE: do not consider {the most dissimilar feature}"*.
- **Ablation A2 - Similar Feature Selection:** The model consistently selects the most similar feature, with the prompt: *"NOTE: you must consider {the most similar feature}"*.

Ablation B - Analysing the Impact of the Hallucination Score Component. This experiment targets scenarios where enrichment was skipped due to the score threshold, testing the hypothesis that enrichment in these cases might impair performance. In this experiment, the model performs only one loop for all questions, analysing instances that would not have received enrichment under the SAFE pipeline.

7 Discussion

The results demonstrate the effectiveness of SAFE in mitigating hallucinations and improving LLM performance. As shown in Tab. 1, SAFE consistently improves accuracy across four diverse datasets. Tab. 4 indicates that enhancing a smaller model with SAFE can yield performance comparable to or

Ablation	TruthfulQA	BioASQ	WikiDoc	HaluEval
Base model	63.63	41.77	38.34	69.47
Ablation A1	46.1 ↓	40.75 ↓	29.3 ↓	51.25 ↓
Ablation A2	61.02 ↓	44.27 ↑	30.48 ↓	63.77 ↓
Ablation B	51.98 ↓	36.0 ↓	32.15 ↓	54.32 ↓

Table 5: Ablation study results on Gemma2-9b. Arrows indicate performance changes relative to the base model (without SAFE).

better than its larger counterpart. The only exception is the LLaMA model on BioASQ and HaluEval, where the 70B variant significantly outperforms the 8B model with SAFE. However, SAFE still provides measurable gains for the smaller model, underscoring its practical utility.

The ablation results in Tab. 5 highlight the contributions of key components. Ablations A1 and A2 show that detecting and removing misleading features improves performance, although simply removing them is insufficient when such features are absent. Conversely, focusing solely on reliable features can overly narrow the model's attention. Ablation B confirms the importance of hallucination score-based uncertainty estimation, as indiscriminate enrichment degrades performance.

Together, these findings demonstrate that SAFE's synergy of hallucination detection and SAE-guided enrichment enhances the reliability of LLMs without requiring additional model training.

8 Conclusion

Hallucination remains a persistent challenge in LLM-based applications, undermining their reliability and trustworthiness in real-world deployments. In this work, we propose SAFE, a Sparse Autoencoder-based Framework for Robust Query Enrichment, which mitigates hallucinations by refining input queries and guiding model responses through interpretable, semantically grounded feature selection. SAFE employs a two-stage process: first, it detects hallucinations using SOTA hallucination detection algorithms; then, it mitigates these issues by enriching queries with features derived from a SAE. Empirical evaluations across diverse benchmark datasets demonstrate that SAFE significantly reduces hallucination rates while improving response accuracy by up to 29.45%. Ablation studies confirm the critical role of detection and SAE-driven enrichment in achieving these gains. Notably, SAFE operates in a training-free manner, offering a lightweight, plug-and-play solution that seamlessly integrates into existing LLM pipelines without additional model fine-tuning.

9 Limitations

While SAFE demonstrates promising results in hallucination mitigation, it has certain limitations. First, the reliance on an SAE and the availability of auto-interpretable features constrain its applicability to LLMs that expose such internal representations. Extending the approach to models without these characteristics would require modifications or alternative interpretability techniques. Second, the effectiveness of the method is inherently influenced by the quality of the input queries. Although this is a common challenge across LLM-based systems, we explicitly acknowledge it here, as low-quality queries may still lead to suboptimal performance. Nevertheless, our evaluation on benchmark datasets, which span diverse query distributions, underscores the robustness and generalizability of our framework. Finally, our current implementation is restricted to English-language inputs, leaving multilingual and multimodal extensions as promising directions for future research.

References

Samir Abdaljalil, Hasan Kurban, Parichit Sharma, Erchin Serpedin, and Rachad Atat. 2025. Sindex: Semantic inconsistency index for hallucination detection in llms. *Preprint*, arXiv:2503.05980.

Kola Ayonrinde, Michael T. Pearce, and Lee Sharkey. 2024. Interpretability as compression: Reconsidering sae explanations of neural activations with mdl-saes. *Preprint*, arXiv:2410.11179.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*.

Trenton Bricken, Jonathan Marcus, Siddharth Mishra-Sharma, Meg Tong, Ethan Perez, Mrinank Sharma, Kelley Rivoire, Thomas Henighan, and Adam Jermyn. 2024. Using dictionary learning features as classifiers. https://transformer-circuits.pub/2024/features-as-classifiers/index.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher

Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Mirko Cesarini, Lorenzo Malandri, Filippo Pallucchini, Andrea Seveso, and Frank Xing. 2024. Explainable ai for text classification: Lessons from a comprehensive evaluation of post hoc methods. *Cognitive Computation*, 16(6):3077–3095.

Zhongzhi Chen, Xingwu Sun, Xianfeng Jiao, Fengzong Lian, Zhanhui Kang, Di Wang, and Chengzhong Xu. 2024. Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20967–20974.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. UPRISE: Universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337, Singapore. Association for Computational Linguistics.

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

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *Preprint*, arXiv:2308.11764.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630. © 2024. The Author(s).

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Davide Ghilardi, Federico Belotti, Marco Molinari, and Jaehyuk Lim. 2024. Accelerating sparse autoencoder training via layer-wise transfer learning in large language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 530–550, Miami, Florida, US. Association for Computational Linguistics.

Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. 2023. Doc2query–: when less is more. In *European Conference on Information Retrieval*, pages 414–422. Springer.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8669–8687, Bangkok, Thailand. Association for Computational Linguistics.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

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

Erik Jones, Hamid Palangi, Clarisse Simões Ribeiro, Varun Chandrasekaran, Subhabrata Mukherjee, Arindam Mitra, Ahmed Hassan Awadallah, and Ece Kamar. 2024. Teaching language models to hallucinate less with synthetic tasks. In *The Twelfth International Conference on Learning Representations*.

Nicola Jones. 2025. Ai hallucinations can't be stopped—but these techniques can limit their damage. *Nature*, 637(8047):778–780.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

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

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023c. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2024. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, Navid Nobani, Andrea Seveso, et al. 2022. The good, the bad, and the explainer: A tool for contrastive explanations of text classifiers. In *IJCAI*, pages 5936–5939.

Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Filippo Pallucchini. 2025. Re-fin: Retrieval-based enrichment for financial data. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 751–759.

Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin T. Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *ArXiv*, abs/2305.15852.

Navid Nobani, Fabio Mercorio, Mario Mezzanzanica, et al. 2021. Towards an explainer-agnostic conversational xai. In *IJCAI*, pages 4909–4910.

Filippo Pallucchini, Xulang Zhang, Rui Mao, and Erik Cambria. 2025. Self-explanatory and retrieval-augmented llms for financial sentiment analysis. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 131–137.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *Preprint*, arXiv:2302.12813.

Daniele Potertì, Andrea Seveso, and Fabio Mercorio. 2025. Can role vectors affect llm behaviour? In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558, Miami, Florida, US. Association for Computational Linguistics.

ZhongXiang Sun, Zihua Si, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, and Jun Xu. 2025. Largepig for hallucination-free query generation: Your large language model is secretly a pointer generator. In *THE WEB CONFERENCE 2025*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall,

Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Bricken Trenton, Jonathan Marcus, Siddharth Mishra-Sharma, Meg Tong, Ethan Perez, Mrinank Sharma, Kelley Rivoire, and Thomas Henighan. 2024. Using dictionary learning features as classifiers. *Transformer Circuits Thread*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weißenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. Simple synthetic data reduces sycophancy in large language models. *Preprint*, arXiv:2308.03958.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2024. Alleviating hallucinations of large language models through induced hallucinations. *Preprint*, arXiv:2312.15710.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.