Novelty Assessment Report

Paper: Hallucination Reduction with CASAL: Contrastive Activation Steering for Amortized Learning

PDF URL: https://openreview.net/pdf?id=YM3RcI3q0E

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Large Language Models (LLMs) exhibit impressive capabilities but often hallucinate, confidently providing incorrect answers instead of admitting ignorance. Prior work has shown that models encode linear representations of their own knowledge and that activation steering can reduce hallucinations. These approaches, however, require real-time monitoring and intervention during inference. We introduce Contrastive Activation Steering for Amortized Learning (CASAL), an efficient algorithm that connects interpretability with amortized optimization. CASAL directly bakes the benefits of activation steering into model's weights. Once trained, LLMs answer questions they know while abstaining from answering those they do not. CASAL's light-weight design requires training only a submodule of a single transformer layer and yet reduces hallucination by $\sim30\%$-40% across multiple short-form QA benchmarks. CASAL is $\sim$30x more compute-efficient and $\sim$20x more data-efficient than strong LoRA-based baselines such as SFT and DPO, boosting its practical applicability in data scarce domains. Importantly, CASAL also generalizes effectively to out-of-distribution (OOD) domains. We showcase CASAL's flexibility in mitigating hallucinations in both text-only and vision-language models. To our knowledge, CASAL is the first steering-based training method that has been shown to be effective for both dense and Mixture-of-Experts (MoE) models. CASAL represents a promising step forward for applying interpretability-inspired method for practical deployment in production systems.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Reducing Hallucinations in Large Language Models through Activation Steering**

A total of **50 papers** were analyzed and organized into a taxonomy with **29 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Activation Steering Methods for Hallucination Mitigation**
- **Vision-Language Models Hallucination Mitigation**
- **Video-Language Models Hallucination Mitigation**
- **Internal State Analysis and Truthfulness Detection**
- **Training-Based Approaches for Hallucination Reduction**
- **Contrastive Decoding and Output Manipulation**
- **Knowledge-Augmented Hallucination Mitigation**
- **Bias and Fairness Analysis in LLMs**
- **Deception and Lying Behavior in LLMs**
- **Hallucination Categorization and Characterization**
- ... and 5 more categories

Complete Taxonomy Tree

- Reducing Hallucinations in Large Language Models through Activation Steering Survey Taxonomy
- Activation Steering Methods for Hallucination Mitigation
  - Contrastive Activation Steering ★ (3 papers)
  - [0] Hallucination Reduction with CASAL: Contrastive Activation Steering for Amortized Learning (Anon et al., 2026) View paper
  - [6] Steering Llama 2 via Contrastive Activation Addition (Schulz, 2023) View paper
  - [21] Differentially Private Steering for Large Language Model Alignment (Goel, 2025) View paper
  - Adaptive and Query-Specific Steering (3 papers)
  - [23] Truthflow: Truthful llm generation via representation flow correction (Wang Hanyu, 2025) View paper
  - [28] Adaptive Activation Steering: A Tuning-Free LLM Truthfulness Improvement Method for Diverse Hallucinations Categories (Tianlong Wang, 2024) View paper
  - [35] QSE: Mitigating LLM Hallucinations Through Query-Adaptive Saliency-Localized Activation Editing (Kewei Liao, 2025) View paper
  - Concept and Representation Space Steering (3 papers)
  - [5] Controlling large language models through concept activation vectors (Hanyu Zhang, 2025) View paper
  - [12] Spectral editing of activations for large language model alignment (Qiu Yifu, 2024) View paper
  - [22] Truthx: Alleviating hallucinations by editing large language models in truthful space (Feng Yang, 2024) View paper
  - Sparse Representation and Feature-Based Control (3 papers)
  - [7] Understanding the Repeat Curse in Large Language Models from a Feature Perspective (Junchi Yao, 2025) View paper
  - [11] Interpretable LLM Guardrails via Sparse Representation Steering (He Zeqing, 2025) View paper
  - [14] Enhancing multiple dimensions of trustworthiness in LLMs via sparse activation control (Xiaofei He, 2024) View paper
  - Gradient-Based Attribution and Steering (2 papers)
  - [40] GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs (Nguyen, 2025) View paper

  - [45] Where Did It Go Wrong? Attributing Undesirable LLM Behaviors via Representation Gradient Tracing (Li Zhe, 2025) View paper
  - Weighted and Controlled Activation Steering (1 papers)
  - [46] Guiding Giants: Lightweight Controllers for Weighted Activation Steering in LLMs (Elhoushi, 2025) View paper
  - Personalized and Behavioral Steering (1 papers)
  - [8] Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization (Zhang Tian-rong, 2024) View paper
- Vision-Language Models Hallucination Mitigation
  - Latent Space Steering for Vision-Language Models (3 papers)
  - [1] Reducing hallucinations in large vision-language models via latent space steering (S Liu, 2025) View paper
  - [2] Reducing Hallucinations in Vision-Language Models via Latent Space Steering (Liu Sheng, 2024) View paper
  - [43] SHARP: Steering Hallucination in LVLMs via Representation Engineering (Junfei Wu, 2025) View paper
  - Visual Information and Token-Level Steering (2 papers)
  - [3] The Hidden Life of Tokens: Reducing Hallucination of Large Vision-Language Models via Visual Information Steering (Li, 2025) View paper
  - [29] Seeing It or Not? Interpretable Vision-aware Latent Steering to Mitigate Object Hallucinations (Chen Boxu, 2025) View paper
  - Activation-Based Decoding for Vision-Language Models (1 papers)
  - [25] Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention (Jingran Su, 2025) View paper
  - Input-Dependent Multimodal Steering (1 papers)
  - [33] Learning to steer: Input-dependent steering for multimodal llms (Parekh, 2025) View paper
  - Sparse Autoencoder-Based Vision-Language Steering (1 papers)
  - [48] Steering LVLMs via Sparse Autoencoder for Hallucination Mitigation (Jinghan He, 2025) View paper
- Video-Language Models Hallucination Mitigation (1 papers)
  - [4] Mitigating Hallucination in VideoLLMs via Temporal-Aware Activation Engineering (Cai Jianfeng, 2025) View paper
- Internal State Analysis and Truthfulness Detection
  - Truthfulness Encoding in Internal States (3 papers)
  - [9] Llms know more than they show: On the intrinsic representation of llm hallucinations (Orgad, 2024) View paper
  - [13] The internal state of an LLM knows when it's lying (Amos Azaria, 2023) View paper
  - [15] Factual Self-Awareness in Language Models: Representation, Robustness, and Scaling (Tamoyan, 2025) View paper
  - Attention and Activation Pattern Analysis (2 papers)
  - [20] UniBias: Unveiling and Mitigating LLM Bias through Internal Attention and FFN Manipulation (Zijian Feng, 2024) View paper
  - [37] Attention Satisfies: A Constraint-Satisfaction Lens on Factual Errors of Language Models (Yuksekgonul, 2023) View paper
  - Sharpness-Based Hallucination Detection (1 papers)
  - [19] In-context sharpness as alerts: An inner representation perspective for hallucination mitigation (Chen Shi-qi, 2024) View paper
  - Factuality Prediction and Confidence Estimation (2 papers)
  - [38] Llm factoscope: Uncovering llms' factual discernment through measuring inner states (Jinwen He, 2024) View paper
  - [41] InternalInspector : Robust Confidence Estimation in LLMs through Internal States (M Beigi, 2024) View paper
- Training-Based Approaches for Hallucination Reduction
  - Self-Evaluation and Self-Alignment (1 papers)
  - [18] Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation (Jin Li-feng, 2024) View paper
  - Supervised Factuality Training (2 papers)
  - [16] Trustworthy reasoning: Evaluating and enhancing factual accuracy in llm intermediate thought processes (Jiao Rui, 2025) View paper
  - [17] The Curious Case of Factuality Finetuning: Models' Internal Beliefs Can Improve Factuality (Newman, 2025) View paper
- Contrastive Decoding and Output Manipulation (1 papers)
  - [10] Mitigating Hallucinations in Large Vision-Language Models with Internal Fact-based Contrastive Decoding (Wang Chao, 2025) View paper
- Knowledge-Augmented Hallucination Mitigation (2 papers)
  - [42] Improving Factuality in LLMs via Inference-Time Knowledge Graph Construction (Wu Shang-Lin, 2025) View paper
  - [50] D-SMART: Enhancing LLM Dialogue Consistency via Dynamic Structured Memory And Reasoning Tree (Lei Xiang, 2025) View paper
- Bias and Fairness Analysis in LLMs (1 papers)
  - [34] Silenced Biases: The Dark Side LLMs Learned to Refuse (Levi, 2025) View paper
- Deception and Lying Behavior in LLMs (2 papers)
  - [36] When Truthful Representations Flip Under Deceptive Instructions? (Xianxuan Long, 2025) View paper
  - [49] Can LLMs Lie? Investigation beyond Hallucination (Prabhudesai, 2025) View paper
- Hallucination Categorization and Characterization (2 papers)
  - [26] Measuring self-deceptive consistency boundaries in large language models through spurious semantic closure networks (A Cranswick, 2025) View paper
  - [44] HACK: Hallucinations Along Certainty and Knowledge Axes (Simhi, 2025) View paper
- Adversarial Robustness and Backdoor Attacks (1 papers)
  - [24] Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models (Pengzhou Cheng, 2024) View paper
- Specialized Hallucination Phenomena (1 papers)
  - [47] Mitigating Misleadingness in LLM-Generated Natural Language Explanations for Recommender Systems: Ensuring Broad Truthfulness Through Factuality and â€¦ (U Maes, 2025) View paper
- Factual Knowledge Limitations and Inconsistencies (2 papers)
  - [32] Semantic and factual alignment for trustworthy large language model outputs (Lian, 2024) View paper
  - [39] Llms as repositories of factual knowledge: Limitations and solutions (Mousavi, 2025) View paper
- Neuron-Level Manipulation and Model Behavior (2 papers)
  - [27] Internal Activation as the Polar Star for Steering Unsafe LLM Behavior (Qian Cheng, 2025) View paper

- [30] Small changes, big impact: How manipulating a few neurons can drastically alter LLM aggression (Jaewook Lee, 2025) View paper

- Misalignment and Safety Surveys (1 papers)
  - [31] Beyond Intentions: A Critical Survey of Misalignment in LLMs. (Yubin Qu, 2025) View paper

Narrative

Core task: Reducing hallucinations in large language models through activation steering. The field has organized itself around several complementary strategies for mitigating hallucinations in large language models. Activation steering methods form a central branch, encompassing techniques that manipulate internal representations to guide model behavior toward truthfulness—ranging from contrastive activation approaches like Contrastive Activation Addition[6] to more sophisticated methods such as Latent Space Steering[1] and Concept Activation Vectors[5]. Parallel branches address modality-specific challenges in vision-language and video-language models, while internal state analysis focuses on detecting and interpreting truthfulness signals within model activations. Training-based approaches and knowledge-augmented methods offer longer-term solutions, and contrastive decoding techniques manipulate outputs at generation time. Additional branches explore neuron-level interventions, adversarial robustness, and broader safety concerns including bias, deception, and misalignment.

Within activation steering, a particularly active line of work centers on contrastive methods that derive steering vectors by comparing activations from truthful versus hallucinated contexts. CASAL[0] exemplifies this approach by learning contrastive activation steering vectors to reduce hallucinations during inference, positioning itself alongside foundational contrastive techniques like Contrastive Activation Addition[6] and more recent innovations such as Internal Contrastive Decoding[10]. These methods share the insight that internal representations encode truthfulness signals that can be amplified or suppressed. Nearby works explore related themes: Hidden Life Tokens[3] investigates how specific token representations influence model behavior, while Sparse Representation Steering[11] and Spectral Activation Editing[12] offer alternative geometric perspectives on activation manipulation. A key tension across these approaches involves balancing intervention strength—steering too aggressively risks degrading fluency or task performance, while subtle interventions may fail to suppress hallucinations reliably. CASAL[0] sits within this dense cluster of contrastive steering methods, emphasizing learned activation adjustments that preserve model capabilities while targeting hallucination-prone representations.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Steering Llama 2 via Contrastive Activation Addition

Authors: Schulz, Julian | **Year/Venue**: 2023 • Annual Meeting of the Association for Computational Linguistics | **URL**: View paper

Abstract

We introduce Contrastive Activation Addition (CAA), an innovative method for steering language models by modifying their activations during forward passes. CAA computes"steering vectors"by averaging the difference in residual stream activations between pairs of positive and negative examples of a particular behavior, such as factual versus hallucinatory responses. During inference, these steering vectors are added at all token positions after the user's prompt with either a positive or negative ...

Relationship Analysis

Both papers belong to the Contrastive Activation Steering category, using difference-in-means between positive and negative examples to compute steering vectors that guide model behavior. The candidate paper (Steering Llama 2 via CAA) focuses on inference-time intervention by adding steering vectors during forward passes to control behaviors like factuality, while the original paper (CASAL) addresses the same hallucination problem but through a training-based approach that amortizes the steering process by baking it into model weights. The key distinction is that CASAL eliminates the need for repeated inference-time steering by training a lightweight subnetwork to approximate the steering solution, achieving greater computational efficiency and data efficiency compared to inference-time methods like the candidate paper.

2. Differentially Private Steering for Large Language Model Alignment

Authors: Goel, Anmol, Hu Yaxi, Anmol Goel, Gurevych, et al. (11 authors total) | **Year/Venue**: 2025 • International Conference on Learning Representations | **URL**: View paper

Abstract

Aligning Large Language Models (LLMs) with human values and away from undesirable behaviors (such as hallucination) has become increasingly important. Recently, steering LLMs towards a desired behavior via activation editing has emerged as an effective method to mitigate harmful generations at inference-time. Activation editing modifies LLM representations by preserving information from positive demonstrations (e.g., truthful) and minimising information from negative demonstrations (e.g., halluc...

Relationship Analysis

Both papers belong to the Contrastive Activation Steering category, using contrastive positive-negative example pairs to compute steering vectors for guiding LLM behavior. While CASAL focuses on amortizing activation steering into model weights during training to reduce hallucinations with high efficiency, the candidate paper addresses privacy concerns by introducing differential privacy guarantees to the steering process, protecting private demonstration data from leakage. The key difference is that CASAL optimizes for computational efficiency and embeds steering into weights, whereas the candidate paper prioritizes privacy preservation during the steering alignment process.

Contributions Analysis

Overall novelty summary. The paper introduces CASAL, a method that bakes activation steering benefits directly into model weights through contrastive learning, enabling models to abstain from answering questions they do not know. This work sits within the 'Contrastive Activation Steering' leaf of the taxonomy, which contains only three papers including CASAL itself. The leaf focuses specifically on methods computing steering vectors from contrastive positive-negative example pairs. This represents a relatively sparse research direction within the broader activation steering landscape, suggesting the specific approach of amortizing steering through weight updates rather than inference-time intervention occupies a less crowded niche.

The taxonomy reveals CASAL's position within a dense ecosystem of activation steering methods. Neighboring leaves include 'Adaptive and Query-Specific Steering' (3 papers), 'Concept and Representation Space Steering' (3 papers), and 'Sparse Representation and Feature-Based Control' (3 papers). The broader 'Activation Steering Methods' branch contains seven distinct approaches, indicating substantial research activity in inference-time interventions. CASAL diverges from these neighbors by moving steering from inference to training time, bridging the gap between the 'Activation Steering Methods' branch and the 'Training-Based Approaches' branch, which focuses on weight updates through fine-tuning and alignment.

Among 29 candidates examined across three contributions, no clearly refutable prior work was identified. Contribution A (CASAL framework) examined 10 candidates with 0 refutable; Contribution B (representation-level training objective) examined 10 candidates with 0 refutable; Contribution C (steering-based training for dense and MoE architectures) examined 9 candidates with 0 refutable. This suggests that within the limited search scope, the specific combination of contrastive activation steering with amortized learning through weight updates appears relatively unexplored. The two sibling papers in the same taxonomy leaf focus on inference-time steering rather than training-time amortization, indicating differentiation even within this narrow research direction.

Based on the limited literature search of 29 candidates, CASAL appears to occupy a distinctive position by combining contrastive steering principles with training-time weight updates. The taxonomy structure shows this bridges two major branches—activation steering and training-based approaches—that typically remain separate. However, the analysis covers top-K semantic matches and does not constitute an exhaustive survey of all possible related work in parameter-efficient fine-tuning, representation learning, or hallucination mitigation more broadly.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: CASAL: Contrastive Activation Steering for Amortized Learning

Description: The authors propose CASAL, a training method that embeds activation steering benefits directly into model weights by training a lightweight subnetwork to approximate steering solutions. This approach reduces hallucinations by teaching models to abstain from answering unknown questions while maintaining performance on known queries.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. ASCD: Attention-Steerable Contrastive Decoding for Reducing Hallucination in MLLM

URL: View paper

Brief Assessment

ASCD[62] focuses on steering cross-modal attention during inference in multimodal models to reduce hallucinations, while CASAL embeds activation steering into model weights during training for text-only and vision-language models. The approaches differ fundamentally in their intervention timing (inference vs. training) and technical mechanisms (attention steering vs. weight modification through representation loss).

2. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization

URL: View paper

Brief Assessment

Personalized Steering Vectors[8] focuses on bi-directional preference optimization for extracting steering vectors to control LLM behavior, while CASAL trains a lightweight subnetwork to embed activation steering into model weights for hallucination reduction. These are distinct technical approaches addressing different problems.

3. Adaptive Activation Steering: A Tuning-Free LLM Truthfulness Improvement Method for Diverse Hallucinations Categories

URL: View paper

Brief Assessment

Adaptive Activation Steering[28] focuses on tuning-free inference-time intervention for truthfulness improvement across diverse hallucination categories, while CASAL embeds activation steering into model weights through training a lightweight subnetwork. These represent fundamentally different approaches—inference-time steering versus training-time weight modification.

4. Hallucination augmented contrastive learning for multimodal large language model

URL: View paper

Brief Assessment

Hallucination Augmented Contrastive[61] focuses on multimodal hallucination reduction through contrastive learning between visual and textual representations, using hallucinative captions as hard negatives. CASAL addresses text-only hallucinations through activation steering embedded in model weights during training, operating on residual stream activations rather than cross-modal alignment.

5. Regularized Contrastive Decoding with Hard Negative Samples for LLM Hallucination Mitigation

URL: View paper

Brief Assessment

Regularized Contrastive Decoding[63] focuses on decoding-time hallucination mitigation through contrastive decoding with hard negative samples, not on training methods that embed activation steering into model weights through amortized optimization.

6. Reducing hallucinations in large vision-language models via latent space steering

URL: View paper

Brief Assessment

Latent Space Steering[1] focuses on reducing hallucinations in vision-language models through visual and textual intervention during inference, while CASAL addresses hallucination reduction in language models through training-time weight updates using contrastive activation steering. The modalities, mechanisms, and intervention timing differ fundamentally.

7. Attention-guided self-reflection for zero-shot hallucination detection in large language models

URL: View paper

Brief Assessment

Attention-guided Self-Reflection[64] focuses on zero-shot hallucination detection using attention mechanisms to guide query reformulation, not on training methods that embed activation steering into model weights for reducing hallucinations.

8. Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention

URL: View paper

Brief Assessment

Activation Steering Decoding[25] focuses on inference-time intervention in vision-language models through bidirectional hidden state manipulation, not on training methods that embed steering into model weights for text-only LLMs.

9. Learning to steer: Input-dependent steering for multimodal llms

URL: View paper

Brief Assessment

Input-dependent Steering[33] focuses on multimodal LLMs with input-specific steering vectors for safety and hallucination tasks, while CASAL addresses text-only models using amortized optimization to embed steering into weights during training. The technical approaches and model architectures differ substantially.

10. Steering Llama 2 via Contrastive Activation Addition

URL: View paper

Brief Assessment

Contrastive Activation Addition[6] focuses on inference-time steering by adding vectors during forward passes, whereas CASAL embeds steering into model weights during training through amortized optimization. The candidate does not demonstrate prior work on training-based amortization of activation steering.

Contribution 2: Representation-level training objective without cross-entropy loss

Description: The method uses a local representation loss applied to residual stream activations as the sole training objective, rather than using it as an auxiliary signal alongside standard cross-entropy loss. This enables efficient single-layer training by providing learning signals from the model's own hidden representations.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Cross-Domain Pre-training with Language Models for Transferable Time Series Representations

URL: View paper

Brief Assessment

Cross-Domain Time Series[56] focuses on time series representation learning using reconstruction objectives for tokenized sequences, not on training language models with representation-level losses as the sole objective for hallucination reduction.

2. Pretraining context compressor for large language models with embedding-based memory

URL: View paper

Brief Assessment

Context Compressor Pretraining[59] focuses on compressing context into embedding representations for efficient LLM inference, not on training language models using only representation-level objectives. The paper uses reconstruction and completion tasks for pretraining a compressor, which is a different technical approach from CASAL's representation loss as the sole training objective.

3. Point Linguist Model: Segment Any Object via Bridged Large 3D-Language Model

URL: View paper

Brief Assessment

Point Linguist Model[58] focuses on bridging representation gaps between LLMs and 3D point clouds for segmentation tasks, not on training language models using only representation-level objectives without cross-entropy loss.

4. Nv-embed: Improved techniques for training llms as generalist embedding models

URL: View paper

Brief Assessment

NV-Embed[53] uses contrastive learning with InfoNCE loss for training embedding models, not representation-level objectives applied to residual stream activations as the sole training signal. The candidate focuses on text embedding tasks rather than hallucination reduction through activation steering.

5. Make large language model a better ranker

URL: View paper

Brief Assessment

Better Ranker[55] focuses on ranking tasks in recommender systems using soft lambda loss combined with cross-entropy for supervised fine-tuning, not on representation-level objectives as the sole training signal for language models.

6. On the role of pretrained language models in general-purpose text embeddings: A survey

URL: View paper

Brief Assessment

Pretrained Embeddings Survey[60] focuses on general-purpose text embeddings using contrastive learning on pairwise datasets, not on training language models with representation-level objectives as the sole training signal without cross-entropy loss.

7. Improving text embeddings with large language models

URL: View paper

Brief Assessment

Text Embeddings LLMs[51] uses standard contrastive loss (InfoNCE) with cross-entropy for training text embeddings, not a representation-level objective. The method focuses on synthetic data generation and fine-tuning with contrastive loss, which is fundamentally different from the original paper's approach of using only local representation loss on residual stream activations.

8. Training Large Language Models to Reason in a Continuous Latent Space

URL: View paper

Brief Assessment

Continuous Latent Reasoning[52] focuses on using continuous hidden states for reasoning in a chain-of-thought paradigm, not on training objectives that replace cross-entropy loss with representation-level losses.

9. Probing the Robustness of Large Language Models Safety to Latent Perturbations

URL: View paper

Brief Assessment

Latent Perturbations Safety[57] focuses on probing safety robustness through latent perturbations and adversarial training, not on replacing cross-entropy loss with representation-level objectives for general language model training.

10. Detoxifying Large Language Models via Autoregressive Reward Guided Representation Editing

URL: View paper

Brief Assessment

Autoregressive Reward Editing[54] focuses on test-time detoxification through representation editing guided by a reward model, not on training language models using only representation-level objectives as the sole training signal.

Contribution 3: First steering-based training framework for both dense and MoE architectures

Description: The authors demonstrate that CASAL is architecture-agnostic and modality-agnostic, successfully reducing hallucinations in both dense transformers and Mixture-of-Experts models, as well as in text-only and vision-language settings.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Mol-MoE: Training Preference-Guided Routers for Molecule Generation

URL: View paper

Brief Assessment

Mol-MoE[66] focuses on mixture-of-experts for molecule generation with preference-based routing, not on steering-based training methods that reduce hallucinations across dense and MoE transformer architectures as in the original paper.

2. Two Experts Are All You Need for Steering Thinking: Reinforcing Cognitive Effort in MoE Reasoning Models Without Additional Training

URL: View paper

Brief Assessment

Two Experts Steering[65] focuses on inference-time steering for MoE reasoning models without training, while the original paper presents a training-based framework that modifies model weights.

3. Steering MoE LLMs via Expert (De)Activation

URL: View paper

Brief Assessment

Expert Activation Steering[69] focuses on inference-time steering of MoE models by detecting and controlling behavior-linked experts, without modifying weights or training. CASAL is a training framework that modifies model weights through amortized optimization. These are fundamentally different approaches (inference-time vs. training-time).

4. MoRE-LLM: Mixture of Rule Experts Guided by a Large Language Model

URL: View paper

Brief Assessment

MoRE-LLM[73] focuses on rule-based interpretability for tabular data using LLM-guided rule extraction, not on steering-based training methods for transformer architectures or hallucination reduction in language models.

5. Steer-MoE: Efficient Audio-Language Alignment with a Mixture-of-Experts Steering Module

URL: View paper

Brief Assessment

Steer-MoE[68] focuses on audio-language alignment using MoE for dynamic steering within audio encoders, not on general steering-based training frameworks across dense and MoE architectures for hallucination reduction in text-only or vision-language models.

6. Steered Mixture-of-Experts for Light Field Images and Video: Representation and Coding

URL: View paper

Brief Assessment

Steered MoE Light[70] focuses on light field image/video representation using mixture-of-experts for compression, not on training frameworks for reducing hallucinations in language models across different architectures.

7. The Compression-Decay Comprehension Test (CDCT): An Information-Theoretic Benchmark for Measuring Machine Comprehension

URL: View paper

Brief Assessment

Compression-Decay Comprehension Test[71] focuses on information-theoretic benchmarking for machine comprehension, not on steering-based training methods for transformer architectures. The candidate does not address training frameworks for dense or MoE models.

8. Multilingual Routing in Mixture-of-Experts

URL: View paper

Brief Assessment

Multilingual Routing MoE[67] focuses on analyzing and intervening in routing patterns within MoE architectures for multilingual tasks, not on steering-based training frameworks that modify model weights during training to reduce hallucinations across both dense and MoE architectures.

9. How to maximize the creativity of artificial intelligence: an experimental analysis of response order and prompting effects

URL: View paper

Brief Assessment

Maximize AI Creativity[72] discusses dense transformer architectures and mixture-of-experts in the context of AI creativity and human steering, not steering-based training methods for hallucination reduction. The technical focus is entirely different from CASAL's contribution.

Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Steering Llama 2 via Contrastive Activation Addition

Detected in: Core Task (sibling), Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Hallucination Reduction with CASAL: Contrastive Activation Steering for Amortized Learning View paper
- [1] Reducing hallucinations in large vision-language models via latent space steering View paper
- [2] Reducing Hallucinations in Vision-Language Models via Latent Space Steering View paper
- [3] The Hidden Life of Tokens: Reducing Hallucination of Large Vision-Language Models via Visual Information Steering View paper
- [4] Mitigating Hallucination in VideoLLMs via Temporal-Aware Activation Engineering View paper
- [5] Controlling large language models through concept activation vectors View paper
- [6] Steering Llama 2 via Contrastive Activation Addition View paper
- [7] Understanding the Repeat Curse in Large Language Models from a Feature Perspective View paper
- [8] Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization View paper
- [9] Llms know more than they show: On the intrinsic representation of llm hallucinations View paper
- [10] Mitigating Hallucinations in Large Vision-Language Models with Internal Fact-based Contrastive Decoding View paper
- [11] Interpretable LLM Guardrails via Sparse Representation Steering View paper
- [12] Spectral editing of activations for large language model alignment View paper
- [13] The internal state of an LLM knows when it's lying View paper
- [14] Enhancing multiple dimensions of trustworthiness in LLMs via sparse activation control View paper
- [15] Factual Self-Awareness in Language Models: Representation, Robustness, and Scaling View paper
- [16] Trustworthy reasoning: Evaluating and enhancing factual accuracy in llm intermediate thought processes View paper
- [17] The Curious Case of Factuality Finetuning: Models' Internal Beliefs Can Improve Factuality View paper
- [18] Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation View paper
- [19] In-context sharpness as alerts: An inner representation perspective for hallucination mitigation View paper
- [20] UniBias: Unveiling and Mitigating LLM Bias through Internal Attention and FFN Manipulation View paper
- [21] Differentially Private Steering for Large Language Model Alignment View paper
- [22] Truthx: Alleviating hallucinations by editing large language models in truthful space View paper
- [23] Truthflow: Truthful llm generation via representation flow correction View paper
- [24] Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models View paper
- [25] Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention View paper
- [26] Measuring self-deceptive consistency boundaries in large language models through spurious semantic closure networks View paper
- [27] Internal Activation as the Polar Star for Steering Unsafe LLM Behavior View paper
- [28] Adaptive Activation Steering: A Tuning-Free LLM Truthfulness Improvement Method for Diverse Hallucinations Categories View paper
- [29] Seeing It or Not? Interpretable Vision-aware Latent Steering to Mitigate Object Hallucinations View paper
- [30] Small changes, big impact: How manipulating a few neurons can drastically alter LLM aggression View paper
- [31] Beyond Intentions: A Critical Survey of Misalignment in LLMs. View paper
- [32] Semantic and factual alignment for trustworthy large language model outputs View paper
- [33] Learning to steer: Input-dependent steering for multimodal llms View paper
- [34] Silenced Biases: The Dark Side LLMs Learned to Refuse View paper
- [35] QSE: Mitigating LLM Hallucinations Through Query-Adaptive Saliency-Localized Activation Editing View paper
- [36] When Truthful Representations Flip Under Deceptive Instructions? View paper
- [37] Attention Satisfies: A Constraint-Satisfaction Lens on Factual Errors of Language Models View paper
- [38] Llm factoscope: Uncovering llms' factual discernment through measuring inner states View paper
- [39] Llms as repositories of factual knowledge: Limitations and solutions View paper
- [40] GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs View paper
- [41] InternalInspector : Robust Confidence Estimation in LLMs through Internal States View paper
- [42] Improving Factuality in LLMs via Inference-Time Knowledge Graph Construction View paper
- [43] SHARP: Steering Hallucination in LVLMs via Representation Engineering View paper
- [44] HACK: Hallucinations Along Certainty and Knowledge Axes View paper
- [45] Where Did It Go Wrong? Attributing Undesirable LLM Behaviors via Representation Gradient Tracing View paper
- [46] Guiding Giants: Lightweight Controllers for Weighted Activation Steering in LLMs View paper
- [47] Mitigating Misleadingness in LLM-Generated Natural Language Explanations for Recommender Systems: Ensuring Broad Truthfulness Through Factuality and â¦ View paper
- [48] Steering LVLMs via Sparse Autoencoder for Hallucination Mitigation View paper
- [49] Can LLMs Lie? Investigation beyond Hallucination View paper
- [50] D-SMART: Enhancing LLM Dialogue Consistency via Dynamic Structured Memory And Reasoning Tree View paper
- [51] Improving text embeddings with large language models View paper
- [52] Training Large Language Models to Reason in a Continuous Latent Space View paper
- [53] Nv-embed: Improved techniques for training llms as generalist embedding models View paper
- [54] Detoxifying Large Language Models via Autoregressive Reward Guided Representation Editing View paper
- [55] Make large language model a better ranker View paper
- [56] Cross-Domain Pre-training with Language Models for Transferable Time Series Representations View paper
- [57] Probing the Robustness of Large Language Models Safety to Latent Perturbations View paper
- [58] Point Linguist Model: Segment Any Object via Bridged Large 3D-Language Model View paper
- [59] Pretraining context compressor for large language models with embedding-based memory View paper
- [60] On the role of pretrained language models in general-purpose text embeddings: A survey View paper

- [61] Hallucination augmented contrastive learning for multimodal large language model View paper
- [62] ASCD: Attention-Steerable Contrastive Decoding for Reducing Hallucination in MLLM View paper
- [63] Regularized Contrastive Decoding with Hard Negative Samples for LLM Hallucination Mitigation View paper
- [64] Attention-guided self-reflection for zero-shot hallucination detection in large language models View paper
- [65] Two Experts Are All You Need for Steering Thinking: Reinforcing Cognitive Effort in MoE Reasoning Models Without Additional Training View paper
- [66] Mol-MoE: Training Preference-Guided Routers for Molecule Generation View paper
- [67] Multilingual Routing in Mixture-of-Experts View paper
- [68] Steer-MoE: Efficient Audio-Language Alignment with a Mixture-of-Experts Steering Module View paper
- [69] Steering MoE LLMs via Expert (De)Activation View paper
- [70] Steered Mixture-of-Experts for Light Field Images and Video: Representation and Coding View paper
- [71] The Compression-Decay Comprehension Test (CDCT): An Information-Theoretic Benchmark for Measuring Machine Comprehension View paper
- [72] How to maximize the creativity of artificial intelligence: an experimental analysis of response order and prompting effects View paper
- [73] MoRE-LLM: Mixture of Rule Experts Guided by a Large Language Model View paper