Hallucinations as Orthogonal Noise: Inference-Time Manifold Alignment via Dynamic Contextual Orthogonalization

Mingkuan Zhao[1]*, Wentao Hu[1,2]*, Tianchen Huang[4], Yuheng Min[5], Suquan Chen[1], Yide Gao[1], Yanbo Zhai[1], Shuangyong Song[2], Xuelong Li[3]†

[1]Xi'an Jiaotong University, [2]Xingchen AGI Lab, China Telecom AI Technology (Beijing) Co., Ltd.,
[3]Institute of Artificial Intelligence, China Telecom, [4]University of Science and Technology of China, [5]Tsinghua University

{mingkuanzhao, wentao_hu, yanbozhai, suquanchen, yidegao}@stu.xjtu.edu.cn,
tchuang@mail.ustc.edu.cn, minyh24@mails.tsinghua.edu.cn,
songshy@chinatelecom.cn, xuelong_li@ieee.org

Abstract

Hallucination in Large Language Models (LLMs)—characterized by the generation of content inconsistent with contextual facts or logical constraints—remains a persistent challenge for reliable deployment. In this work, we address this issue through a geometric framework rooted in the linear representation hypothesis. We propose that hallucinations manifest as *orthogonal noise* relative to the semantic manifold of the residual stream. Specifically, we hypothesize that while attention heads ideally propagate information congruent with the context subspace, hallucinations arise when specific heads introduce components orthogonal to this subspace, disrupting the coherence of the latent representation.

Based on this formulation, we introduce **Dynamic Contextual Orthogonalization (DCO)**, an inference-time intervention method. DCO utilizes the input residual stream as a dynamic context anchor to perform orthogonal decomposition on attention head outputs. To distinguish between context-aligned semantic updates and divergent noise, DCO employs a layer-wise Z-score suppression mechanism that selectively attenuates outlier orthogonal components based on statistical distributions.

Evaluations on Llama-3-8B and 70B across benchmarks such as XSum, NQ-Swap, and IFEval demonstrate that DCO achieves superior contextual faithfulness compared to state-of-the-art intervention baselines. Furthermore, DCO maintains high performance on knowledge-intensive tasks like TriviaQA and TruthfulQA, effectively mitigating the trade-off between hallucination suppression and parametric knowledge retention often observed in existing methods. Our findings validate the geometric interpretation of hallucinations and establish DCO as a computationally efficient approach for enforcing manifold alignment. Our code is available at https://github.com/Harry-Miral/DCO.

1 Introduction

The deployment of Large Language Models (LLMs) in critical domains is hindered by the persistence of hallucinations, where generated content diverges from contextual facts or logical constraints. While current mitigation strategies such as Retrieval-Augmented Generation (RAG) or contrastive decoding effectively modulate output probabilities, they often operate without a mechanistic view of the internal representation dynamics. We posit that hallucinations are not merely surface-level statistical anomalies but manifestations of geometric misalignment within the model's latent space, specifically when the model processes conflicting or ambiguous contexts.

This study approaches the hallucination problem through the *Linear Representation Hypothesis* (Park et al., 2024), which suggests that semantic concepts are encoded as linear directions in high-dimensional space (Gurnee and Tegmark, 2024). Under this framework, we formalize the faithfulness of generation as a manifold alignment task. Ideally, attention heads should propagate information that lies parallel to the subspace defined by the context. We hypothesize that hallucinations arise from *Orthogonal Noise Injection*—a phenomenon where specific attention heads introduce components orthogonal to the established context manifold, thereby steering the latent state evolution away from the factual trajectory.

Based on this geometric formulation, we introduce **Dynamic Contextual Orthogonalization (DCO)**, an inference-time intervention method. Unlike static steering vectors that apply a fixed direction, DCO utilizes the input residual stream of each layer as a dynamic *context anchor*. The method performs orthogonal decomposition on attention head outputs to isolate components that deviate

*These authors contributed equally to this work.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 36575–36586

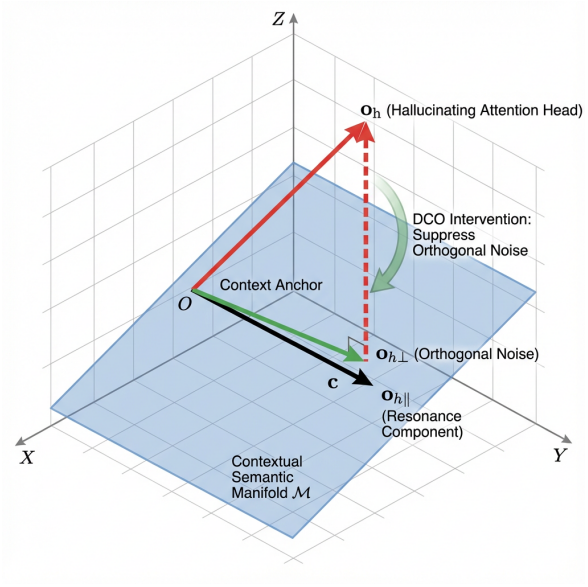

Figure 1: **Conceptual geometry of Dynamic Contextual Orthogonalization (DCO).** A hallucinating attention head produces an output vector $\mathbf{o}_h$ that diverges from the contextual semantic manifold $\mathcal{M}$. DCO decomposes $\mathbf{o}_h$ into a context-aligned resonance component $\mathbf{o}_{h\parallel}$ and an orthogonal noise component $\mathbf{o}_{h\perp}$. By dynamically attenuating $\mathbf{o}_{h\perp}$ based on layer-wise statistical distributions, DCO realigns the latent representation with the established manifold.

from the global semantic consensus. Crucially, to distinguish between necessary information diversification and harmful noise, DCO employs a layer-wise Z-Score suppression mechanism. This adaptive filtering is derived from the observation that hallucinatory signals manifest as statistical outliers in the orthogonality distribution, necessitating a dynamic rather than static threshold to preserve the model's parametric knowledge retrieval capabilities.

We evaluate DCO on the Llama-3-8B and 70B models across a diverse set of benchmarks. Empirical results indicate that DCO achieves superior performance compared to state-of-the-art intervention methods, including Inference-Time Intervention (ITI) and Decoding by Contrasting Layers (DoLa). Specifically, DCO significantly enhances contextual faithfulness on XSum and IFEval while maintaining robustness on knowledge-intensive tasks such as TriviaQA and MuSiQue. These findings demonstrate that enforcing geometric constraints in the residual stream offers a computationally efficient and mechanistically grounded approach to hallucination mitigation, achieving an optimal trade-off between faithfulness and general capability.

2 Related Work

Mechanistic Interpretability of Hallucinations. Understanding the computational underpinnings of Transformer models is essential for diagnosing generation failures. Grounded in the view of the residual stream as a communication channel modulated by independent attention heads (Dar et al., 2023), mechanistic interpretability has identified specific components governing information flow. Key findings include the characterization of induction heads (Olsson et al., 2022) and "retrieval heads" that extract factual content from long contexts (Wu et al., 2024). Moreover, research on knowledge localization (Meng et al., 2022; Geva et al., 2023) elucidates how factual associations are retrieved via the interplay between MLP layers and attention mechanisms. Recent work by Nanda et al. (2023) and Yu et al. (2024) further distinguishes between early-layer knowledge gaps and upper-layer retrieval failures as distinct causes of hallucination. Complementary to this, Zhao et al. (2025b) demonstrate that attention heads exhibit heterogeneous importance, showing that sparse attention patterns can be enforced without a speed-performance trade-off—an observation that supports our hypothesis of differential head behavior in the context of hallucinatory noise. Our work operationalizes these insights, translating the theoretical understanding of attention head dynamics into a direct intervention method within the residual stream.

Inference-Time Hallucination Mitigation. Various training-free decoding strategies have been developed to intervene during inference. Methods such as Contrastive Decoding (CD) (Li et al., 2023) and its variants, including DoLa (Chuang et al., 2023) and Autocontrastive Decoding (Gera et al., 2023), adjust next-token probabilities by contrasting distributions across model layers or scales. Addressing contextual faithfulness specifically, Shi et al. (2024) proposed Context-Aware Decoding (CAD) to amplify evidence-based logits, while Chen et al. (2024) utilized in-context sharpness as a detection signal. More recently, Gema et al. (2024) introduced DeCoRe to mask retrieval heads based on static importance scores. Crucially, these approaches predominantly operate at the output logit level or employ static interventions **(as illustrated in Figure 2)**. In contrast, guided by the *Linear Representation Hypothesis* (Park et al., 2024; Her-

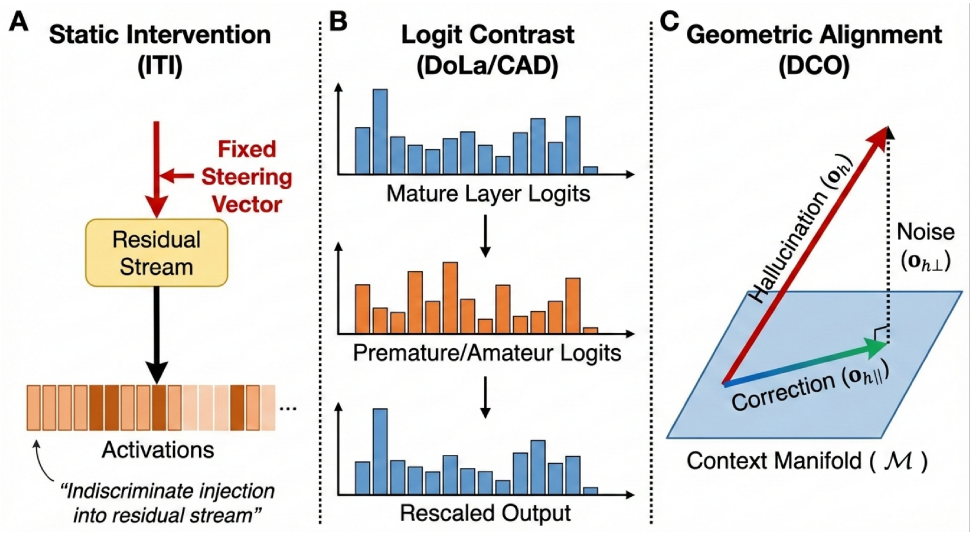

Figure 2: **Mechanistic comparison of intervention paradigms. (A) Static Intervention (ITI):** Injects a fixed steering vector derived from offline probing, applying indiscriminate directionality to the residual stream regardless of context. **(B) Logit Contrast (DoLa/CAD):** Operates at the output layer by subtracting premature/amateur logits from mature layer logits, enhancing sharpness but failing to correct internal representation dynamics. **(C) Geometric Alignment (DCO, Ours):** Performs dynamic intervention within the latent space. It treats hallucination as an orthogonal noise component $\mathbf{o}_{h\perp}$ diverging from the context manifold $\mathcal{M}$ and projects the attention output $\mathbf{o}_h$ to align with the valid semantic subspace $\mathbf{o}_{h\parallel}$.

nandez et al., 2024), DCO implements a *dynamic* orthogonalization mechanism. Unlike Inference-Time Intervention (ITI) (Li et al., 2024a), which applies fixed steering vectors derived from offline probing, DCO dynamically calibrates the intervention based on the instantaneous geometry of the residual stream at each step.

Geometric Perspectives on Semantic Consistency. We conceptualize generative faithfulness as a manifold alignment problem. In contexts involving knowledge conflicts or long-range dependencies, parametric memory often interferes with contextual extraction (Longpre et al., 2021; Liu et al., 2024). While Representation Engineering (Zou et al., 2025) demonstrates the utility of top-down linear control, the geometric interpretation of hallucinations as orthogonal interference remains under-explored. DCO addresses this by reformulating factual consistency—evaluated via metrics such as FactKB (Feng et al., 2023)—as the filtering of orthogonal noise. By integrating challenges identified in summarization benchmarks like XSum (Narayan et al., 2018) into this framework, we provide a unified geometric approach that preserves model integrity while suppressing noise, offering a mechanistic alternative to probability-based corrections.

Large Language Model Development and Downstream Applications. The rapid scaling of LLMs, as exemplified by bilingual model series such as TeleChat (He et al., 2024; Wang et al., 2024) and their successors TeleChat2 and TeleChat2.5 (Wang et al., 2025), as well as the Mixture-of-Experts architecture TeleChat3-MoE (Liu et al., 2025), highlights the increasing complexity and diversity of deployed language systems. Similarly, the Tele-FLM series (Li et al., 2024c,b) demonstrates the empirical lessons learned when scaling from tens to hundreds of billions of parameters. Across these architectures, hallucination persists as a critical reliability bottleneck. This challenge is particularly pronounced in high-stakes downstream applications, including mathematical reasoning (Zhao et al., 2025a), structured table reasoning (Xiong et al., 2025), process-reward-guided structured data construction (Xing et al., 2025), and universal information extraction via reinforcement learning (Li et al., 2025), where factual deviations carry direct consequences. Furthermore, model compression research such as Mosaic Pruning for Mixture-of-Experts models (Hu et al., 2025) underscores the need to maintain representational integrity under efficiency constraints—a concern closely aligned with hallucination-free generation. Our work addresses this fundamental reliability gap through a training-free geometric intervention,

providing a principled solution broadly applicable across the aforementioned deployment contexts.

3 Methodology

The computational architecture of the DCO intervention is illustrated in **Figure 3**. The mechanism of Dynamic Contextual Orthogonalization (DCO) functions by applying linear projection operators to constrain the state evolution of latent representations during inference. This operation targets the output of the Multi-Head Attention (MHA) module in layer L, where vectors produced by individual attention heads are orthogonally decomposed to identify and attenuate components that exhibit statistically significant deviation from the contextual semantic manifold.

3.1 Context Anchor Construction

The initial phase involves constructing a context anchor, $\mathbf{c}$, which quantitatively defines the direction of semantic consistency at the current logical step. We posit that the input residual stream $\mathbf{x}_{in}^{L} \in \mathbb{R}^{d_{model}}$ at layer L represents the accumulated semantic consensus derived from the context processed up to layer $L-1$. To focus on the directional alignment rather than magnitude, we apply Root Mean Square Normalization (RMSNorm) to the residual vector. The normalized context anchor is defined as:

$$\mathbf{c} = \frac{\text{RMSNorm}(\mathbf{x}_{in}^{L})}{\|\text{RMSNorm}(\mathbf{x}_{in}^{L})\|_2 + \epsilon} \tag{1}$$

where ϵ is a small stability constant. Geometrically, $\mathbf{c}$ serves as the reference unit vector on the hypersphere of the latent space, delineating the principal direction of the ongoing generative process.

3.2 Subspace-based Orthogonal Decomposition

The second phase executes a fine-grained decomposition of attention head outputs. Let H denote the number of attention heads in layer L. For the h-th attention head, let $\mathbf{o}_h \in \mathbb{R}^{d_{model}}$ be its output vector. Following the principles of linear projection, we partition $\mathbf{o}_h$ into a **context-aligned component** $\mathbf{o}_{h\parallel}$ and a **context-orthogonal component** $\mathbf{o}_{h\perp}$.

The context-aligned component, representing information congruent with the context anchor, is computed via projection:

$$\mathbf{o}_{h\parallel} = (\mathbf{o}_h \cdot \mathbf{c}) \cdot \mathbf{c} \tag{2}$$

The orthogonal component is defined as the vector rejection:

$$\mathbf{o}_{h\perp} = \mathbf{o}_h - \mathbf{o}_{h\parallel} \tag{3}$$

To quantify the divergence of each head relative to the context, we define the orthogonality metric ρ_h as the ratio of the orthogonal component's magnitude to the total magnitude:

$$\rho_h = \frac{\|\mathbf{o}_{h\perp}\|_2}{\|\mathbf{o}_h\|_2 + \epsilon} \tag{4}$$

As $\rho_h \to 0$, the head reinforces the existing context; as $\rho_h \to 1$, the head introduces information linearly independent of the established semantic manifold.

3.3 Z-Score Based Dynamic Suppression

To achieve adaptive intervention, DCO relies on the statistical distribution of ρ within the current layer rather than static thresholds. We observe that in high-dimensional semantic spaces, valid information diversification and hallucinatory noise often overlap in absolute magnitude but differ in their relative statistical deviation.

We first calculate the mean μ_ρ and standard deviation σ_ρ across all H heads:

$$\mu_\rho = \frac{1}{H}\sum_{h=1}^{H}\rho_h, \quad \sigma_\rho = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(\rho_h - \mu_\rho)^2} \tag{5}$$

The standard score (Z-Score) z_h for each head is derived to measure relative deviation:

$$z_h = \frac{\rho_h - \mu_\rho}{\sigma_\rho + \epsilon} \tag{6}$$

As illustrated in Figure 4, we hypothesize that hallucinating heads manifest as statistical outliers in the long tail of the orthogonality distribution. To softly attenuate these outliers while preserving gradient continuity, we employ a non-linear Sigmoid gating function to determine the suppression coefficient $\lambda_h \in (0, 1]$:

$$\lambda_h = \frac{1}{1 + \exp(\beta \cdot (z_h - \tau))} \tag{7}$$

Here, τ represents the tolerance threshold in standard deviation units, and β is the temperature coefficient controlling the steepness of the suppression curve. The final reconstructed output of the attention module at layer L is the summation of the intervened head vectors:

$$\text{AttentionOutput}_L = \sum_{h=1}^{H}(\mathbf{o}_{h\parallel} + \lambda_h \cdot \mathbf{o}_{h\perp}) \tag{8}$$

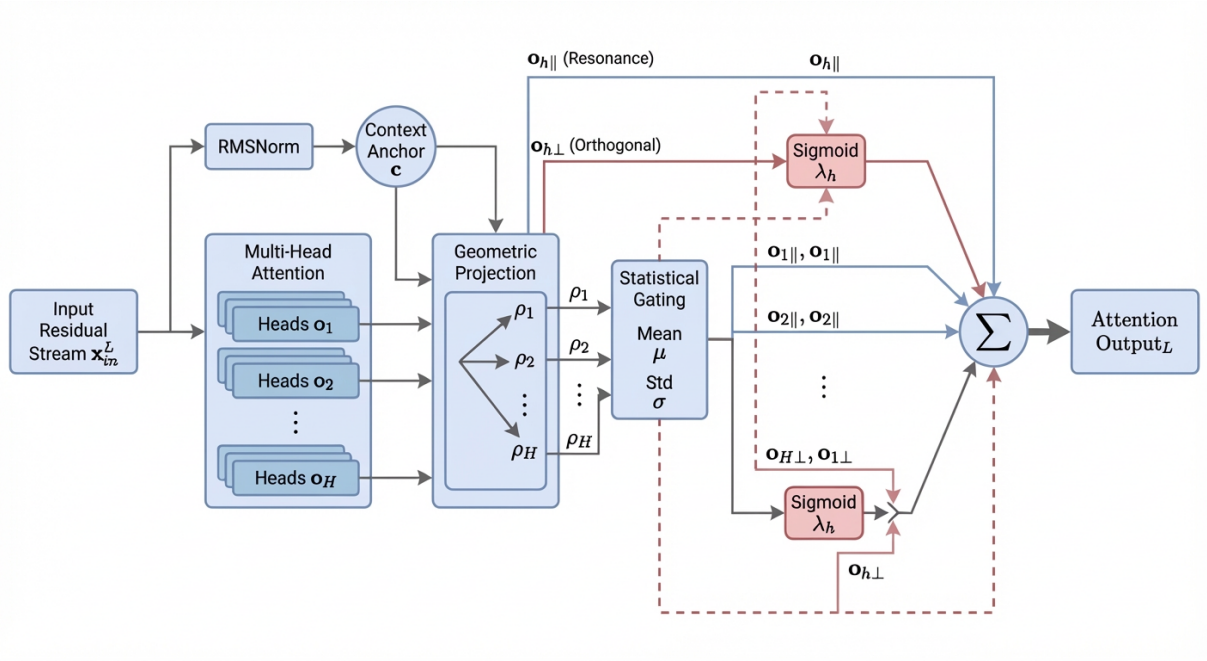

Figure 3: **Mechanistic pipeline of Dynamic Contextual Orthogonalization (DCO).** The process begins by deriving a context anchor $\mathbf{c}$ from the input residual stream $\mathbf{x}_{in}^L$. Each attention head output $\mathbf{o}_h$ is then orthogonally decomposed into a **context-aligned component** $\mathbf{o}_{h\|}$ and an orthogonal component $\mathbf{o}_{h\perp}$. The aligned path bypasses intervention to preserve logical continuity, while the orthogonal path is modulated by a dynamic suppression coefficient λ_h, computed via layer-wise Z-score statistics (μ, σ) of the orthogonality metric ρ. This ensures that only outlier heads—introducing divergent noise relative to the semantic manifold—are attenuated.

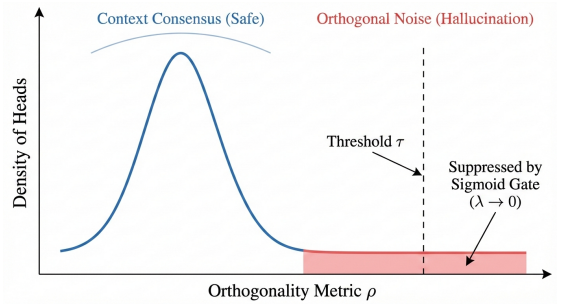

Figure 4: **Schematic of the Z-Score dynamic suppression mechanism.** We model the distribution of orthogonality metrics ρ across attention heads. Most heads cluster within the context-aligned region (blue peak), while hallucinating heads manifest as statistical outliers in the high-orthogonality tail (red region). DCO dynamically sets a threshold τ derived from layer-wise statistics to selectively suppress these divergent components via a Sigmoid gate.

This formulation ensures that heads exhibiting normative behavior ($z_h < \tau$) retain their orthogonal contributions ($\lambda_h \approx 1$), thereby preserving necessary semantic diversity.

3.4 Implementation Strategy

DCO is deployed using a layer-selective strategy, targeting the middle-to-late layers defined by $L \in [L_{start}, L_{end}]$. This design is grounded in mechanistic findings that early layers ($L < L_{start}$) are critical for parsing local syntax and often exhibit high intrinsic orthogonality required for structural composition.

The computational overhead of DCO is negligible. For a model with hidden dimension d_{model} and H heads, the complexity is $O(L \cdot H \cdot d_{model})$. Since d_{model} is constant and $H \cdot d_{model}$ scales linearly with the model size, this operation is significantly more efficient than the quadratic attention mechanism ($O(N^2 \cdot d_{model})$), ensuring minimal latency increase in production environments.

4 Results and Analysis

We evaluate the efficacy of Dynamic Contextual Orthogonalization (DCO) across a spectrum of benchmarks to assess its impact on faithfulness, factuality, and reasoning stability. All experiments are conducted using the Llama-3-8B-Instruct and Llama-3-70B-Instruct models (Grattafiori et al., 2024).

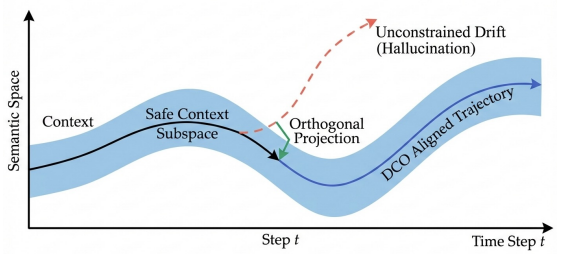

Figure 5: **Visualization of latent manifold dynamics.** The blue ribbon represents the *Context Manifold* $\mathcal{M}$. Without intervention, the generative trajectory (Red dashed line) suffers from *orthogonal drift*, diverging into hallucination. DCO (Blue solid line) applies continuous orthogonal projection (Green arrow) at each step t, constraining the latent state evolution within the safe semantic subspace.

We evaluate contextual faithfulness and instruction following using IFEval (Zhou et al., 2023), XSum, and NQ-Swap. For factuality and knowledge retention, we employ TruthfulQA (Lin et al., 2022), TriviaQA (Joshi et al., 2017), and Natural Questions (NQ-Open) (Kwiatkowski et al., 2019). Complex reasoning stability is assessed via MuSiQue (Trivedi et al., 2022).

4.1 Experimental Configuration

To maintain intervention stability across diverse semantic landscapes, we adopt a calibration strategy based on *semantic density* and *reasoning depth*. The intervention is primarily applied to the middle-to-late layers of the residual stream, aligning with mechanistic findings that these layers govern factual retrieval and integration.

The hyperparameters—tolerance threshold τ and temperature coefficient β—are configured to reflect the geometric constraints of specific tasks. For counterfactual tasks requiring strict adherence to provided evidence (e.g., NQ-Swap), we apply tighter projection constraints to filter parametric bias. Conversely, for multi-hop reasoning (e.g., MuSiQue), constraints are relaxed to accommodate the necessary semantic expansion required for logic chaining. This task-dependent configuration ensures that DCO adapts to the specific manifold geometry of downstream applications.

4.2 Performance on Contextual Faithfulness

Table 1 details the performance of DCO on benchmarks evaluating adherence to external contexts and instructions. DCO demonstrates statistically significant improvements over baselines, particularly in instruction following and context-grounded retrieval tasks.

On the IFEval benchmark, DCO achieves state-of-the-art results across model scales. For Llama-3-70B, DCO elevates the Prompt-level strict accuracy to 85.77% and Instruction-level accuracy to 90.29%, representing a substantial margin over the greedy baseline of 77.45% and 84.41%, respectively, and exceeding existing intervention methods such as DeCoRe. Similarly, in the NQ-Open Open-Book setting, DCO outperforms all baselines, reaching 79.51% on the 70B model. These results suggest that enforcing geometric alignment within the residual stream effectively enhances the capacity of the model to prioritize contextual evidence over internal parametric noise.

Regarding the XSum summarization task, DCO achieves the highest ROUGE-L of 20.39 on the 8B model. Crucially, when compared to Inference-Time Intervention (ITI), which utilizes static steering vectors, DCO exhibits superior robustness. While ITI suffers a severe performance regression, dropping to 13.25 R-L, DCO maintains high linguistic quality. This disparity validates the hypothesis that dynamic, context-dependent orthogonalization is essential for maintaining the integrity of the latent manifold.

In the NQ-Swap counterfactual task, while Context-Aware Decoding (CAD) retains the highest EM scores through logit-level contrastive amplification, DCO serves as a computationally efficient alternative. By achieving competitive alignment of 75.33% EM on the 70B model via a single-forward projection, DCO avoids the dual-decoding latency overhead inherent in contrastive methods while still effectively suppressing parametric interference.

4.3 Pareto Frontier: Knowledge Preservation

An essential consideration in hallucination mitigation is the preservation of the model's general capabilities, as aggressive intervention often leads to a degradation of parametric knowledge. Table 2 illustrates this performance trade-off. While methods such as ITI and CD report higher scores on TruthfulQA, they frequently cause a significant performance regression in factual retrieval. For instance, on the Llama-3-8B model, ITI results in a decrease in TriviaQA accuracy from 56.58% to 48.41%, indicating that static interventions may inadvertently suppress valid factual retrieval pathways alongside hallucinatory signals.

In contrast, DCO optimizes the Pareto fron-

Table 1: Performance comparison on faithfulness evaluation tasks. We report ROUGE-L (R-L), BERTScore (Zhang et al., 2020), and FactKB (Feng et al., 2023) scores. Note that **NQ-Open** here is evaluated in an **Open-Book** setting (with retrieved context).

Model	Method	XSum			MemoTrap		IFEval		NQ-Open	NQ-Swap
		R-L	BS-F1	FactKB	Macro	Micro	Prompt	Inst.	(Open)	EM
Llama-3-8B	Greedy	19.90	67.23	47.61	65.86	64.40	70.24	78.30	69.68	60.62
	ITI	13.25	59.96	34.35	62.65	58.96	52.31	63.19	56.16	51.08
	CAD	18.82	67.20	**67.16**	-	-	-	-	69.83	**74.21**
	DoLa (low)	19.82	67.19	47.21	65.27	63.69	69.69	78.18	69.68	60.77
	DoLa (high)	19.92	67.34	48.49	64.85	63.17	70.24	78.66	69.49	60.98
	AD	19.79	67.31	48.49	65.38	64.28	67.65	76.26	68.93	60.51
	DeCoRe (static)	19.87	67.83	64.07	**69.53**	**69.20**	69.13	78.06	70.62	64.43
	DCO (Ours)	**20.39**	**67.86**	56.70	67.06	65.30	**74.68**	**81.41**	**74.95**	62.96
Llama-3-70B	Greedy	22.41	69.77	61.32	68.47	66.52	77.45	84.41	71.07	76.11
	ITI	21.64	69.46	61.33	71.24	68.73	76.71	83.69	71.90	74.76
	CD	**22.71**	**69.99**	54.73	69.27	67.55	71.72	79.74	65.80	68.37
	CAD	21.45	69.28	**65.61**	-	-	-	-	71.83	**84.70**
	DoLa (low)	22.46	69.80	61.11	67.99	65.93	77.08	84.29	71.07	75.98
	DoLa (high)	22.43	69.93	59.99	67.92	65.81	78.00	84.65	70.40	75.26
	AD	22.49	69.91	60.57	67.51	66.44	76.89	84.41	71.15	74.02
	DeCoRe (static)	21.94	69.35	64.88	**71.96**	**71.41**	78.56	84.89	72.51	79.06
	DCO (Ours)	22.40	69.74	60.87	69.17	67.19	**85.77**	**90.29**	**79.51**	75.33

Table 2: Performance comparison on factuality benchmarks. We focus on TruthfulQA (MC), TriviaQA, and NQ-Open in the closed-book setting.

Model	Method	TruthfulQA (MC)			TriviaQA	NQ-Open
		MC1	MC2	MC3	EM	(Closed)
Llama-3-8B	Greedy	39.41	55.69	30.31	56.58	29.04
	ITI	**43.70**	**62.78**	**34.91**	48.41	22.07
	DoLa (low)	39.05	55.65	30.06	56.63	29.15
	DoLa (high)	38.68	55.64	30.19	56.50	29.19
	AD	31.21	55.30	28.28	54.93	28.32
	DeCoRe (static)	38.68	55.74	29.80	56.93	29.42
	DCO (Ours)	40.39	58.68	32.21	**57.33**	**30.92**
Llama-3-70B	Greedy	49.57	70.60	37.85	74.77	40.08
	ITI	48.96	67.04	37.27	73.54	38.57
	CD	**57.77**	**76.65**	**47.08**	72.83	36.23
	DoLa (low)	49.45	70.58	37.75	74.74	40.08
	DoLa (high)	49.69	70.88	38.01	73.96	39.59
	AD	42.23	67.56	35.37	74.14	40.23
	DeCoRe (static)	51.29	72.02	40.24	74.79	40.41
	DCO (Ours)	48.23	69.56	36.73	**76.12**	**43.62**

tier by maintaining or enhancing performance on knowledge-intensive benchmarks. On the Llama-3-70B model, DCO achieves the highest performance among the evaluated methods for both TriviaQA (76.12%) and NQ-Open Closed-book (43.62%) settings. This robustness is attributed to the adaptive Z-score statistical filtering mechanism. By dynamically defining suppression thresholds based on layer-wise distributions, DCO distinguishes between the normative statistical signals of factual retrieval and the outlier components characteristic of hallucinations. These results provide empirical evidence that dynamic orthogonalization effectively minimizes interference with standard knowledge retrieval processes while maintaining competitive truthfulness scores, such as the 40.39% MC1 achieved on the 8B model.

4.4 Reasoning Stability in Complex Chains

Table 3 evaluates the performance of various intervention methods on the MuSiQue multi-hop reasoning dataset. On the Llama-3-70B model, DCO achieves the highest performance in the Closed-book setting when utilized with Chain-of-Thought (CoT), reaching an Exact Match (EM) score of 21.14%. This represents an improvement over both the greedy baseline (20.15%) and head-level intervention methods such as DeCoRe (20.60%).

The observed improvement is posited to stem from the mitigation of compounding deviations within the residual stream during multi-step inference. In complex reasoning chains, minor orthogonal components introduced in initial logical steps may accumulate, leading to semantic drift where the latent trajectory diverges from the factual man-

Table 3: Multi-hop reasoning performance on MuSiQue.

Model	Method	Without CoT		With CoT	
		Closed Book	Open Book	Closed Book	Open Book
Llama-3-8B	Greedy	7.41	58.83	14.61	69.84
	ITI	4.01	45.84	4.18	38.31
	CAD	-	57.88	-	**73.02**
	DoLa	7.24	59.08	**14.94**	69.92
	AD	6.99	58.63	14.40	69.92
	DeCoRe (static)	**7.90**	**61.23**	14.69	72.49
	DCO (Ours)	7.41	58.26	14.56	69.01
Llama-3-70B	Greedy	11.79	68.56	20.15	74.43
	ITI	10.88	68.14	20.44	74.27
	CD	10.92	66.61	17.17	71.70
	CAD	-	68.64	-	74.02
	DoLa	11.42	68.68	20.15	74.64
	AD	11.38	68.14	20.23	74.27
	DeCoRe (static)	11.79	**69.76**	20.60	**75.05**
	DCO (Ours)	**11.96**	67.89	**21.14**	74.27

ifold. By enforcing manifold alignment at each layer, DCO attenuates these divergent components before they propagate through the network. Furthermore, the competitive performance of DCO relative to head-masking approaches like DeCoRe suggests that the errors associated with multi-hop reasoning may be distributive in nature. Consequently, these errors appear to be more effectively addressed via global statistical constraints on the residual stream rather than the isolation of individual attention heads. While DoLa and DeCoRe maintain a slight advantage in certain Open-book scenarios, DCO demonstrates robust stability in internalizing reasoning paths without external evidence.

4.5 Qualitative Analysis

To examine the microscopic mechanism of DCO, we analyze a "knowledge conflict" scenario in Figure 6, where fictitious context contradicts the model's parametric priors. In the baseline model, specific attention heads dominated by pre-trained associations generate real-world entities (e.g., "Apple"), effectively ignoring the provided context. Geometrically, these heads produce output vectors that are linearly independent of the context anchor $\mathbf{c}$, representing orthogonal drift from the semantic manifold.

DCO identifies these divergent components as statistical outliers in the Z-score distribution of the orthogonality metric ρ. By applying the dynamic gating mechanism, DCO attenuates these anomalous orthogonal vectors while preserving components collinear with the context anchor. This operation effectively projects the divergent generative trajectory back onto the context manifold $\mathcal{M}$ without disrupting syntactic coherence. These findings provide empirical evidence for the hypothesis of hallucinations as orthogonal noise and demonstrate DCO's efficacy in enforcing contextual alignment under conflict.

5 Conclusion

In this work, we have presented Dynamic Contextual Orthogonalization (DCO), a framework that operationalizes the geometric hypothesis of hallucinations as *orthogonal noise injection*. By formalizing factual deviations as vector components that diverge from the contextual semantic manifold, we established a mechanistic basis for intervention. DCO utilizes the instantaneous residual stream as a dynamic anchor and employs a layer-wise Z-Score suppression mechanism. This design addresses the rigidity of static steering vectors by adapting to the local statistical distribution of attention outputs, thereby distinguishing between necessary semantic diversification and anomalous noise.

Empirical validation across the Llama-3 family confirms that DCO effectively resolves the persistent trade-off between contextual faithfulness and parametric knowledge retention. Unlike traditional decoding strategies that often degrade general reasoning capabilities to enforce context adherence, DCO's outlier-based attenuation preserves the model's underlying logical structure while significantly enhancing performance on faithfulness benchmarks such as XSum and IFEval. The method's robustness on TriviaQA and MuSiQue further validates that enforcing geometric constraints is a non-destructive pathway to reli-

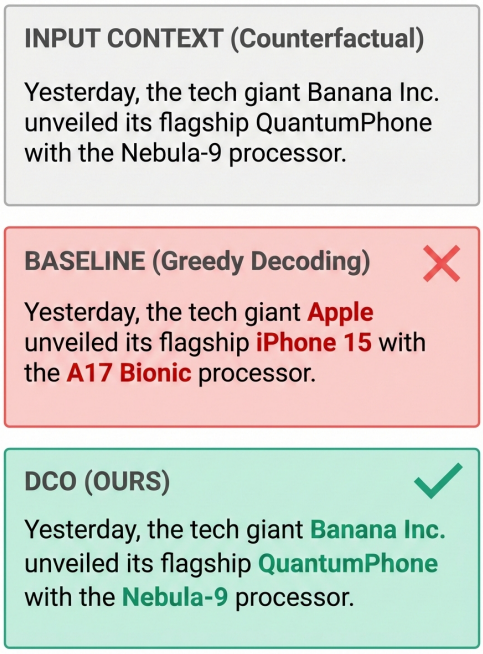

Figure 6: **Qualitative comparison under knowledge conflict.** Top: The input context contains counterfactual entities ("Banana Inc."). Middle: The baseline model (Red) ignores the context and hallucinates real-world entities ("Apple", "iPhone") due to parametric memory bias. Bottom: DCO (Green) successfully suppresses this orthogonal noise, aligning the generation with the counterfactual context.

able generation.

In summary, DCO offers a computationally efficient, training-free mechanism for alignment during inference. It underscores the necessity of dynamic, representation-level intervention for suppressing hallucinations in complex semantic spaces. Future research will focus on developing meta-learning frameworks to automate the calibration of geometric constraints, aiming to achieve self-adaptive manifold control that generalizes across varying semantic densities.

6 Limitations

While Dynamic Contextual Orthogonalization (DCO) demonstrates significant efficacy in enhancing faithfulness and mitigating hallucinations, we identify specific boundary conditions and design choices inherent to the framework.

The implementation of DCO introduces additional linear projection and statistical normalization operations at each decoding step. While this theoretically increases the floating-point operations (FLOPs) per layer, the added complexity scales linearly with the model dimension ($O(d_{model})$). This is negligible compared to the quadratic complexity of the self-attention mechanism ($O(N^2)$) inherent to Transformer architectures. Furthermore, unlike contrastive decoding paradigms (e.g., CAD, DoLa) which necessitate multiple forward passes or parallel decoding of different model states to compute logit differences, DCO operates strictly within a single inference pass. Consequently, in production environments prioritizing throughput and latency, DCO maintains a distinct efficiency advantage over multi-pass intervention strategies.

The geometric definition of the semantic manifold in DCO relies on the extraction of a context anchor from the input residual stream. This design is explicitly optimized for context-grounded generation, summarization, and context-grounded reasoning, where an external reference exists. In scenarios entirely devoid of context (pure closed-book generation), the definition of orthogonality becomes less distinct. However, our empirical results on knowledge-intensive benchmarks (e.g., TriviaQA, TruthfulQA) indicate that the dynamic Z-score filtering mechanism is robust enough to distinguish between valid parametric retrieval and stochastic noise even in lower-context settings. This suggests that DCO successfully mitigates the trade-off between hallucination suppression and general capability degradation, a common failure mode in static intervention methods.

Our framework is predicated on the Linear Representation Hypothesis, positing that semantic deviations manifest as orthogonal vectors in the residual stream. While the underlying neural architecture involves non-linear activation functions, the effectiveness of DCO supports the utility of linear approximations for manipulating high-level semantic features. This geometric simplification provides a mechanistically interpretable control method, offering a more transparent alternative to black-box modifications of output probability distributions.

References

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada. Association for Computational Linguistics.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952, Singapore. Association for Computational Linguistics.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2024. Decore: Decoding by contrasting retrieval heads to mitigate hallucinations. *arXiv preprint arXiv:2410.18860*.

Ariel Gera, Roni Friedman, Ofir Arviv, Chulaka Gunasekara, Benjamin Sznajder, Noam Slonim, and Eyal Shnarch. 2023. The benefits of bad advice: Autocontrastive decoding across model layers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10406–10420, Toronto, Canada. Association for Computational Linguistics.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

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

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. *Preprint*, arXiv:2310.02207.

Zhongjiang He, Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, Yan Wang, Xin Wang, Luwen Pu, Huinan Xu, Ruiyu Fang, Yu Zhao, Jie Zhang, Xiaomeng Huang, Zhilong Lu, and 17 others. 2024. Telechat technical report. *Preprint*, arXiv:2401.03804.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. *Preprint*, arXiv:2308.09124.

Wentao Hu, Mingkuan Zhao, Shuangyong Song, Xiaoyan Zhu, Xin Lai, and Jiayin Wang. 2025. Mosaic pruning: A hierarchical framework for generalizable pruning of mixture-of-experts models. *Preprint*, arXiv:2511.19822.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

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

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024a. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Chao Wang, Xinzhang Liu, Zihan Wang, Yu Zhao, Xin Wang, Yuyao Huang, Shuangyong Song, Yongxiang Li, Zheng Zhang, Bo Zhao, Aixin Sun, Yequan Wang, Zhongjiang He, Zhongyuan Wang, Xuelong Li, and Tiejun Huang. 2024b. 52b to 1t: Lessons learned via tele-flm series. *Preprint*, arXiv:2407.02783.

Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Chao Wang, Xinzhang Liu, Zihan Wang, Yu Zhao, Xin Wang, Yuyao Huang, Shuangyong Song, Yongxiang Li, Zheng Zhang, Bo Zhao, Aixin Sun, Yequan Wang, Zhongjiang He, Zhongyuan Wang, Xuelong Li, and Tiejun Huang. 2024c. Tele-flm technical report. *Preprint*, arXiv:2404.16645.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Zhongqiu Li, Shiquan Wang, Ruiyu Fang, Mengjiao Bao, Zhenhe Wu, Shuangyong Song, Yongxiang Li, and Zhongjiang He. 2025. Mr-uie: Multi-perspective reasoning with reinforcement learning for universal information extraction. *Preprint*, arXiv:2509.09082.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xinzhang Liu, Chao Wang, Zhihao Yang, Zhuo Jiang, Xuncheng Zhao, Haoran Wang, Lei Li, Dongdong He, Luobin Liu, Kaizhe Yuan, Han Gao, Zihan Wang, Yitong Yao, Sishi Xiong, Wenmin Deng, Haowei He, Kaidong Yu, Yu Zhao, Ruiyu Fang, and 35 others. 2025. Training report of telechat3-moe. *Preprint*, arXiv:2512.24157.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *Preprint*, arXiv:2301.05217.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Zihan Wang, XinZhang Liu, Shixuan Liu, Yitong Yao, Yunyao Huang, Mengxiang Li, Zhongjiang He, Yongxian Li, Luwen Pu, Huinan Xu, Chao Wang, and Shuangyong Song. 2024. TeleChat: An open-source billingual large language model. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 10–20, Bangkok, Thailand. Association for Computational Linguistics.

Zihan Wang, Xinzhang Liu, Yitong Yao, Chao Wang, Yu Zhao, Zhihao Yang, Wenmin Deng, Kaipeng Jia, Jiaxin Peng, Yuyao Huang, Sishi Xiong, Zhuo Jiang, Kaidong Yu, Xiaohui Hu, Fubei Yao, Ruiyu Fang, Zhuoru Jiang, Ruiting Song, Qiyi Xie, and 19 others. 2025. Technical report of telechat2, telechat2.5 and t1. *Preprint*, arXiv:2507.18013.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *Preprint*, arXiv:2404.15574.

Hongrui Xing, Xinzhang Liu, Zhuo Jiang, Zhihao Yang, Yitong Yao, Zihan Wang, Wenmin Deng, Chao Wang, Shuangyong Song, Wang Yang, Zhongjiang He, and Yongxiang Li. 2025. LLMSR@XLLM25: A language model-based pipeline for structured reasoning data construction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 342–350, Vienna, Austria. Association for Computational Linguistics.

Sishi Xiong, Dakai Wang, Yu Zhao, Jie Zhang, Changzai Pan, Haowei He, Xiangyu Li, Wenhan Chang, Zhongjiang He, Shuangyong Song, and Yongxiang Li. 2025. Tablereasoner: Advancing table reasoning framework with large language models. *Preprint*, arXiv:2507.08046.

Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. 2024. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Deji Zhao, Donghong Han, Jia Wu, Zhongjiang He, Bo Ning, Ye Yuan, Yongxiang Li, Chao Wang, and Shuangyong Song. 2025a. Enhancing math reasoning ability of large language models via computation logic graphs. *Knowledge-Based Systems*, 325:113905.

Mingkuan Zhao, Wentao Hu, Jiayin Wang, Xin Lai, Tianchen Huang, Yuheng Min, Rui Yan, and Xiaoyan Zhu. 2025b. Making every head count: Sparse attention without the speed-performance trade-off. *Preprint*, arXiv:2511.09596.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and

Le Hou. 2023. Instruction-following evaluation for large language models. *ArXiv*, abs/2311.07911.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.