How Context Shapes Truth: Geometric Transformations of Statement-level Truth Representations in LLMs

Shivam Adarsh
University of Copenhagen
shad@di.ku.dk

Maria Maistro
University of Copenhagen
mm@di.ku.dk

Christina Lioma
University of Copenhagen
c.lioma@di.ku.dk

Abstract

Large Language Models (LLMs) often encode whether a statement is true as a vector in their residual stream activations. These vectors, also known as *truth vectors*, have been studied in prior work, however how they change when context is introduced remains unexplored. We study this question by measuring (1) the directional change (θ) between the truth vectors with and without context and (2) the relative magnitude of the truth vectors upon adding context. Across four LLMs and four datasets, we find that (1) truth vectors are roughly orthogonal in early layers, converge in middle layers, and may stabilize or continue increasing in later layers; (2) adding context generally increases the truth vector magnitude, i.e., the separation between true and false representations in the activation space is amplified; (3) larger models distinguish relevant from irrelevant context mainly through directional change (θ), while smaller models show this distinction through magnitude differences. We also find that context conflicting with parametric knowledge produces larger geometric changes than parametrically aligned context. Collectively, these findings provide a geometric characterization of how context transforms the truth vector in the activation space of LLMs.[1]

1 Introduction

As Large Language Models (LLMs) get increasingly adopted in high stakes applications, it becomes important to understand how they process and represent information internally. Prior work (Hollinsworth et al., 2024; Gurnee and Tegmark, 2023; Marks and Tegmark, 2024) studies how concepts are encoded in model activations, specifically using activations from residual stream (after the MLP layer).[2] They find that many high-level concepts, including whether a statement is true, are represented as linear directions (i.e. vectors) in the activation space (termed "truth directions"). Prior work (Burns et al., 2023; Azaria and Mitchell, 2023; Marks and Tegmark, 2024; Li et al., 2023; Bao et al., 2025) shows that linear classifiers can reliably separate true from false statements in the LLM activation space, implying a geometric structure to how truth is represented. However, these studies do not study how this geometry changes when context is added. While in-context learning

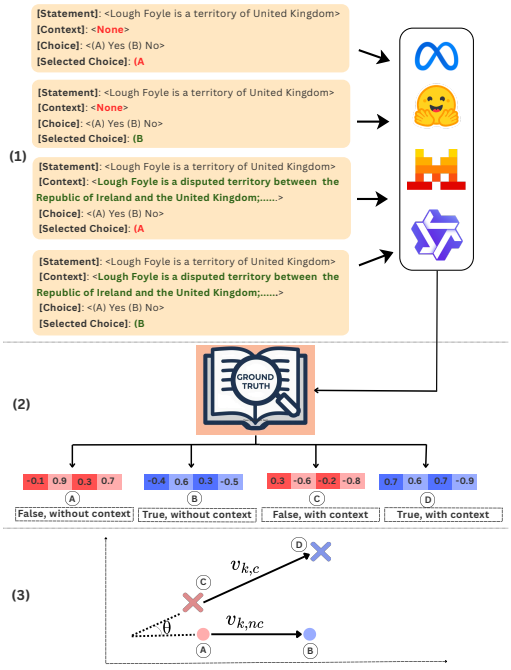


Figure 1: **Overview of our approach (1)** For a statement k, we generate 4 inputs by varying the [Selected Choice] and presence of context. The LLM is instructed to generate the completion based on the [Selected Choice]. **(2)** We extract the residual stream activations for generating the first token and label them as true or false based on the ground truth. **(3)** We compare the truth vectors with and without context ($v_{k,nc}$ and $v_{k,c}$), calculating directional change θ and relative magnitudinal change $\frac{||v_{k,c}||^2}{||v_{k,nc}||^2}$ across all the layers.

[1] Our code is available here

[2] In the rest of the paper, by "residual stream" we will mean after the MLP layer without explicitly clarifying it.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36570–36590
July 2-7, 2026 ©2026 Association for Computational Linguistics

and retrieval-augmented-generation have proven effective at improving model outputs without retraining (Brown et al., 2020; Min et al., 2022; Wei et al., 2023; Lewis et al., 2020; Gao et al., 2023), how the geometric structure of statement-level truth changes when context is added remains underexplored. It is precisely these geometric changes in the direction and magnitude of residual stream activations when context is added that we study in this work. We contribute the first characterisation of how truth geometry transforms when context is added. Understanding this has theoretical implications for how LLMs process context, and practical implications for designing retrieval-augmented and in-context learning systems that more reliably integrate contextual knowledge.

We analyze the residual stream activations when an LLM processes a statement with and without context. For both conditions, we extract the vectors in activation space that separate true from false statements i.e., the "*truth vectors*". We hypothesize that adding context should alter this geometric structure. To test this, we examine two geometric properties: the angle between the truth vectors with and without context (θ), which captures directional change, and the relative magnitude of the truth vectors, which captures whether context amplifies or compresses the separation between true and false representations in the activation space. Figure 1 gives an overview of our approach.

Experiments with four LLMs and four datasets, spanning diverse domains and context types, show the following three findings: (1) **Three-phase pattern of directional change**: Comparing truth vectors with and without context, we find that truth vectors are approximately orthogonal in early layers, converge sharply in early to middle layers, and then either stabilize or continue increasing in later layers depending on the dataset. (2) **Increase in Relative Magnitudes**: Adding context generally increases the truth vector magnitude, i.e., the separation between true and false representations in the activation space increases. (3) **Sensitivity to relevant vs irrelevant context**: On comparing relevant context with randomly generated and irrelevant context, we find that relevant context generally produces a higher directional or magnitudinal change. These findings are statistically significant across models and datasets. Collectively, our results provide novel empirical evidence on how context reshapes the geometric structure of statement-level truth representations in the LLM's activation space.

2 Related Work

Truth Representations in LLMs Understanding how LLMs represent truth has received attention. Burns et al. (2023) introduce Contrast-Consistent Search (CCS), an unsupervised methodology showing that truth directions can be extracted from model activations. This work shows that LLMs encode truth as a linear direction in their representation space. Marks and Tegmark (2024) extend this using mass-mean probes, which compute the mean difference between activations for true and false statements to identify truth directions. Li et al. (2023) introduce Inference-Time Intervention (ITI), showing that shifting model activations along truthful directions can significantly improve LLM truthfulness. This work distinguishes between generation accuracy (measured by model output) and probe accuracy (classifying statements using intermediate activations); similarly to this, our work also focuses on internal representations rather than output behavior. Bürger et al. (2024) address the failure of truth probes to generalize across negated statements by showing that truth is represented in a two-dimensional subspace rather than a single direction. Lastly, Bao et al. (2025) find that consistent truth directions emerge in more capable models and that probes trained on factual statements generalize to in-context settings, including question answering grounded in provided passages and abstractive summarization. However, they test whether a single probe transfers across these settings, not whether the geometric structure of truth vectors change when context is introduced. Our work addresses this gap directly.

While the above work establishes that truth has a geometric structure in the activation space of LLMs and tests if truth probes generalize in different settings, it does not directly examine how truth vectors change when context is added. It is precisely this gap that our work addresses by measuring the geometric transformations, namely, the directional change θ and relative magnitude shift between truth vectors with and without context, showing that context induces consistent layer-dependent changes.

Activation Steering and Contrastive Vectors Prior work has shown that LLM behavior can be steered by adding contrastive vectors to model activations (Turner et al., 2024; Rimsky et al., 2024; Zou et al., 2023; Subramani et al., 2022). These vectors are typically computed as the mean difference between the activations of two contrasting

conditions, such as truth and false (Li et al., 2023), toxic and non-toxic (Liu et al., 2024), or positive and negative sentiment (Turner et al., 2024). During inference, the contrasting vectors are added back to shift the model's behavior. Here, magnitude of the contrastive vector is a key hyperparameter, serving as the strength of intervention. We take inspiration from steering techniques for our method and instead of using vectors to modify behavior, we observe how truth vectors change when context is introduced.

Context Utilization Research on in-context learning has focused on how models use instructions and exemplars to recognize tasks and learn input-output mappings (Brown et al., 2020; Min et al., 2022; Wei et al., 2023), with evidence that task recognition occurs in the middle layers (Sia et al., 2024). While prior work has focused on how LLMs utilize context by analyzing the generated outputs (Du et al., 2024; Marjanovic et al., 2024; Hagström et al., 2025) or by probing the residual stream activations directly to detect knowledge conflict signals (Zhao et al., 2024), our work focuses on how context geometrically changes the direction of truth in the residual stream activations.

3 Methodology

Task Description To analyze how context induces changes to the truth vector, we set up a text generation task as shown in Figure 2. Given a statement, we create four prompts: supporting or refuting the statement, each with or without context. The LLM is instructed to continue the generation supporting or refuting the statement. Figure 2a shows the complete prompt for generating a completion that supports the statement, with context. We ensure that the first token generated by the model is ")" for a fair comparison between generations supporting and refuting the statements. We vary the [Selected Choice] field in the prompt (Figure 2a) to generate the completions and randomize the choices within the [Choice] field to remove any bias arising from the ordering of the options. This setting is similar to that described in Rimsky et al. (2024). Next, we use the ground truth labels from the dataset to map generations into true and false.

Truth Vectors from Residual Stream Activations We define the truth vector by taking the difference of the residual stream activations used to produce the first token in [completion] for true and false

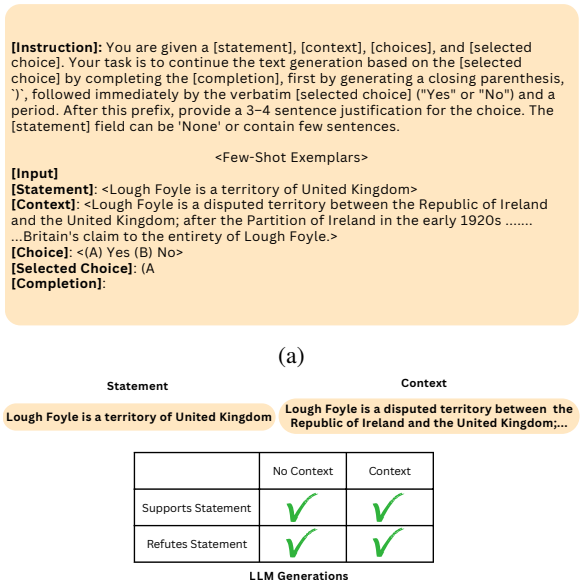


Figure 2: (a) Prompt to generate completion supporting the statement with context. (b) Given a statement, we create four prompts: supporting or refuting the statement, each with or without context. The LLM generates a completion for each prompt.

generations in each layer. We extract the activations used to generate the first output token (i.e., activations at the final token position of the prompt) as this position aggregates information from the entire input via causal attention and is not influenced by subsequently generated tokens (Marks and Tegmark, 2024; Burns et al., 2023). For a statement k, we define the truth vector in layer l as:

$$v_k^{(l)} = a_{k,\text{True}}^{(l)} - a_{k,\text{False}}^{(l)} \quad (1)$$

where $a_{k,\text{True}}^{(l)}$ (resp. $a_{\text{k,False}}^{(l)}$) are the residual stream activations to generate the first token for true (resp. false) completions. We extend Eq. 1 to account for the generation with and without context:

$$v_{k,\text{nc}}^{(l)} = a_{k,\text{True, nc}}^{(l)} - a_{k,\text{False, nc}}^{(l)} \quad (2)$$

$$v_{k,\text{c}}^{(l)} = a_{k,\text{True, c}}^{(l)} - a_{k,\text{False, c}}^{(l)} \quad (3)$$

For a statement k, $v_{k,\text{nc}}^{(l)}$ (resp. $v_{k,\text{c}}^{(l)}$) is the truth vector in layer l without context (resp. with context).

Calculating Theta and Relative Magnitude We compute how adding context changes the truth vector through two geometric properties: directional change θ and relative magnitude. θ is the directional change between the truth representations with and without context. A large θ implies that the

directions of truth are fundamentally different in the residual stream. For a statement k, we compute the directional change θ in layer l as:

$$\theta_k^{(l)} = \arccos\left(\frac{v_{k,c}^{(l)}.v_{k,nc}^{(l)}}{||v_{k,c}^{(l)}||.||v_{k,nc}^{(l)}||}\right) \quad (4)$$

For a dataset D, we average across all the statements N_k to get the θ in layer l as:

$$\theta_D^{(l)} = \frac{\sum_k \theta_k^{(l)}}{|N_k|} \quad (5)$$

where $|N_k|$ is the total number of statements.

Relative magnitude signifies the separation between the true and false representations in the residual stream. Values above 1 mean that context increases the separation between true and false representations, while values below 1 mean that context decreases the separation. To calculate relative magnitudes, we use the L_2 norm distance between true and false representations when no context is present as a baseline (see AB in Figure 3). Next, we check if the distance between true and false representations increases or decreases when context is added. For a statement k, we compute the increase in relative magnitudes between true and false representations when context is added as:

$$rm_{k,tc-fc}^{(l)} = \frac{||v_{k,c}^{(l)}||^2}{||v_{k,nc}^{(l)}||^2} \quad (6)$$

Eq. 6 corresponds to measuring $\frac{CD}{AB}$ in Figure 3. For a dataset D, we average across all statements N_k to get the relative magnitudes for the entire dataset in each layer l as:

$$rm_{D,tc-fc}^{(l)} = \frac{1}{|N_k|}\sum_k rm_{k,tc-fc}^{(l)} \quad (7)$$

where $|N_k|$ is the total number of statements. We also calculate the vectors $v_{k,tc-fnc}^{(l)}$ and $v_{k,tnc-fc}^{(l)}$ to measure the relative magnitudes in the case when context is added to generate either true or false completions while generating the other completion without any context (see Appendix A.2).

4 Experimental Set-up

We aim to study how the geometric structure of the truth vector (specifically its directional change θ (Eq. 5) and relative magnitude (Eq. 7)) changes when context is introduced. We select only statements where the LLM follows instructions across all four prompts (Figure 2b); see Appendix A.8.

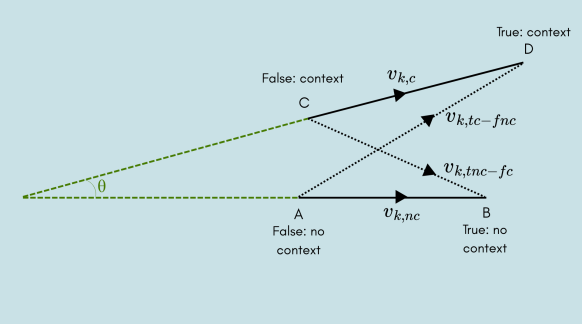


Figure 3: For a statement k, $v_{k,nc}$ (AB) is the truth vector without context and $v_{k,c}$ (CD) is the truth vector when context is added. θ is the angle between $v_{k,nc}$ and $v_{k,c}$ denoting the directional change (Eq. 4). To track relative magnitudes, we compute the ratio of L_2 distances: $\frac{CD}{AB}$, $\frac{AD}{AB}$ and $\frac{BC}{AB}$ as per Eq. 6, 10 & 11.

LLMs We use four instruction-tuned models spanning different scales (3B–12B) and families: LLama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-Nemo-12B-Instruct (Mistral AI and NVIDIA, 2024), Qwen3-4B-Instruct (Yang et al., 2025) and SmolLM3-3B (Bakouch et al., 2025). This selection allows us to examine whether the observed geometric transformations generalize across model scale. As our task is text-generation following a specific set of instructions, we use off-the-shelf instruction fine-tuned models. We use Huggingface API for inference with greedy decoding sampling to ensure reproducibility. All experiments were conducted on NVIDIA A100 and H100 GPUs, requiring approximately 500 GPU hours.

Datasets We use datasets containing statements and relevant contexts: Druid (Hagström et al., 2025), MF2 (Zaranis et al., 2025), ConflictQA (Xie et al., 2024) and LegalBench (Guha et al., 2023). We select three subsets from Druid: Borderlines, Politifact and ScienceFeedback and analyze them separately as the context type varies across them. See Table 1 for a dataset summary and Appendix A.1 for details. While Druid, MF2 and LegalBench contain real world data, ConflictQA is a synthetic dataset. We use two subsets from ConflictQA: Parametric and Counter. ConflictQA-Parametric contains context which is aligned to the LLM's parametric knowledge and ConflictQA-Counter contains context which goes against the parametric knowledge. In Table 1 we also show the Fleisch score of context, which approximates human difficulty in understanding text (Flesch, 1948).

Dataset	Rows	Len.	Read.	Context Type
Borderlines	982	153.9	41.3	Geo. factcheck
Politifact	907	114.6	47.7	Pol. factcheck
ScienceFeedback	618	128.5	39.6	Sci. factcheck
MF2	1736	457.9	57.2	Movie synopsis
CL-Bill	500	185.3	6.9	Legal bills
CL-Company	500	657.5	10.7	Company descr.
ConflictQA-Counter	1244	82.4	46.0	Parametrically counter context
ConflictQA-Parametric	1244	50.8	53.9	Parametrically aligned context

Table 1: Dataset statistics. Len. is mean context length in words. Read. is the Flesch Reading Ease (0–100, the lower, the harder the text). Borderlines, Politifact and ScienceFeedback are subsets of DRUID. CL denotes Corporate Lobbying datasets from LegalBench.

5 Experiments and Discussion

5.1 Directional Change across Layers

To understand how context changes the truth representations in the residual stream, we begin with a layer-wise analysis of directional change θ. Figure 4 shows how θ changes across layers. A lower θ means higher similarity between truth vectors with and without context. All four LLMs show a consistent 3-phase pattern: θ remains high (near orthogonal) in early layers, drops sharply in middle layers to reach a minimum, and then either stabilizes or increases in later layers. LLaMA and Mistral begin decreasing around layer 9, reaching minima near layer 15, while smaller models (Qwen, SmolLM) show prolonged early phases until layers 14–16 with later minima (layers 20–25). In later layers, behavior varies by model-dataset combination. This 3-phase pattern, and especially the convergence of the truth vectors in the middle layers, is consistent with prior findings that early layers handle low-level input processing, middle layers encode semantic information, and later layers shift toward next-token prediction (Ghandeharioun et al., 2024). Early LLM layers have been related to capturing syntactic meaning (Li and Subramani, 2025). As such, the direction of "truth" can potentially have less meaning in early layers - leading to orthogonality in phase-1. We also verify this using probes built to classify truth using residual stream activations. We observe that accuracies often peak in the middle layers and are usually low in the earlier layers (Appendix A.4). Further, we note that while larger models compress the initial stage into fewer layers (until layer 9), smaller models take longer (until layers 14-16).

The convergence in the middle layers indicates that truth vectors with and without context become more similar. This aligns with prior work showing that middle layers are the primary site for semantic encoding (Ghandeharioun et al., 2024; Geva et al., 2023), factual knowledge retrieval (Meng et al., 2022), and task-relevant representations (Hendel et al., 2023). Ghandeharioun et al. (2024) observe that steering vectors are most effective in middle layers, where input processing has concluded but next-token prediction has not yet dominated, and Sia et al. (2024) find that task recognition in machine translation occurs in similar layers. Similarly to Azaria and Mitchell (2023), we also observe that accuracies of probes often peak in the middle layers (Appendix A.4). Notably, θ never reaches zero, suggesting that while truth vectors converge, the models maintain distinct representations for statements with and without context.

In phase 3, θ shows a flat trend for most datasets, suggesting that truth vectors have largely converged by the middle layers. However, for ConflictQA-Counter and Politifact, θ increases in later layers for LLaMA, Mistral, and Qwen, possibly reflecting continued processing of context that conflicts with parametric knowledge. Notably, θ values for ConflictQA-Counter consistently exceed those for ConflictQA-Parametric, indicating that contradictory context induces greater directional shift than aligned context. This is consistent with prior findings that LLMs exhibit confirmation bias towards memory-aligned information (Xie et al., 2024) and that knowledge conflicts arise from competing memory heads and context heads in later layers (Jin et al., 2024). These findings suggest that when context aligns with parametric knowledge, both pathways reinforce the same truth direction and θ stabilizes, whereas when context contradicts it, competing signals persist through later layers, producing continued divergence. Further, prior work suggests that deeper layers are often redundant and can be pruned with limited performance loss (Men et al., 2025), though the final layer remains important. Our results suggest that later-layer contributions may also be context-dependent. We further verify that these truth vectors are causally functional through interventional experiments where steering along these directions reliably flips model outputs (Appendix A.9).

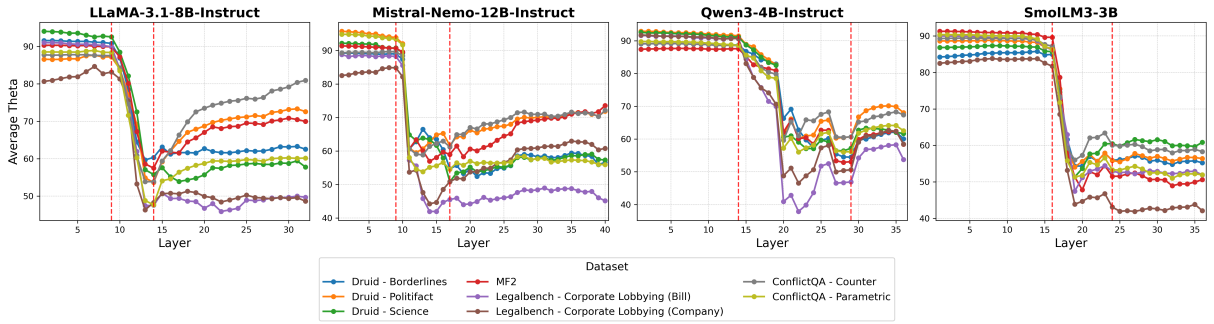


Figure 4: Layer wise plot of average θ in degrees across different models and datasets indicating the directional change in truth vectors when context is added. Vertical red lines indicate the beginning of a new phase. Across all the settings, we observe three phases: Phase-1, where the truth vectors are almost orthogonal, Phase-2, where the truth vectors become more similar and finally Phase-3, where truth vectors stabilize or continue increasing.

5.2 Relative Magnitude

To understand how context affects the separation between true and false representations, we compute the relative magnitude of the truth vector when context is added (as described in Section 3. The results for the relative magnitudes in the final layer are shown in Table 2 and a layerwise analysis is presented in Figure 5. Relative magnitude values A above (resp. below) 1 mean that the separation between true and false representations increase (resp. decrease) when context is added.

Figure 5 shows a certain variability across LLM layers with respect to relative magnitude, however one common pattern is a peak in middle layers followed by a decline and eventual stabilization. LLaMA shows early-layer variability, an upward spike around layers 15–20, then stabilization. Mistral exhibits a spike around layers 10–15, a sharp decline through layers 15–20, then stabilization. Qwen shows a spike around layer 22, then declining until layer 27 before stabilizing. SmolLM displays a spike around layers 17–19, a slight decrease, and a secondary smaller spike around layers 25–27, though this later spike is absent for MF2 and Corporate Lobbying. Across models, middle-layer spikes almost always exceed 1, indicating that the separation between true and false representations are maximum in the middle layers. Notably, middle layers are often responsible for semantic encoding (Ghandeharioun et al., 2024). Although relative magnitudes decrease toward later layers, they generally remain above 1, even in the final layers (Table 2). In the final layer, LLaMA increases the average relative magnitude of the truth vector across 7 out of 8 datasets. However, the results are mixed for other models. Additional results are found in Appendix A.2 and Appendix A.7.

Dataset	LLaMA	Mistral	Qwen	SmolLM
Borderlines	**1.18***	**1.08***	**1.13***	**1.11***
Politifact	1.01	0.85	**1.07***	**1.15***
ScienceFeedback	**1.10***	0.87	1.00	**1.19***
MF2	**1.13***	1.00	**1.06***	0.96
CL-Bill	**1.06***	**1.06***	1.00	0.95
CL-Company	**1.15***	**1.18***	**1.06***	1.00
ConflictQA - Counter	**1.20***	0.98	0.98	**1.26***
ConflictQA - Param	**1.34***	1.02	**1.06***	**1.16***

Table 2: Relative magnitude (Eq. 7) averaged over statements from the final LLM layer across datasets. Values above 1 mean that the truth vector magnitude increases when context is added. * marks stat. significance of $p < 0.05$ with the Wilcoxon signed-rank test.

Note that we also examine whether θ and relative magnitude correlate with changes in output probability for “True" and “False" tokens when context is added (Appendix A.5). We find some correlations, but not consistently across datasets and models. This suggests that θ and relative magnitudes do not necessarily translate to probabilistic differences in the output generation.

5.3 Relevant versus Random Context

Motivated by prior work showing that adding unrelated context dramatically reduces model performance (Shi et al., 2023; Yoran et al., 2024), we compare the effect of adding relevant versus random context to study if relevant context produces different geometric changes than random context, we experiment with five different contexts varying in degree of randomness: (1) context of “random characters”, such that words have no linguistic meaning; (2) context of “random words”, randomly sampled from the NLTK english corpus and ordered randomly, such that the sentence has no meaning; (3) context of “random salad”, where the

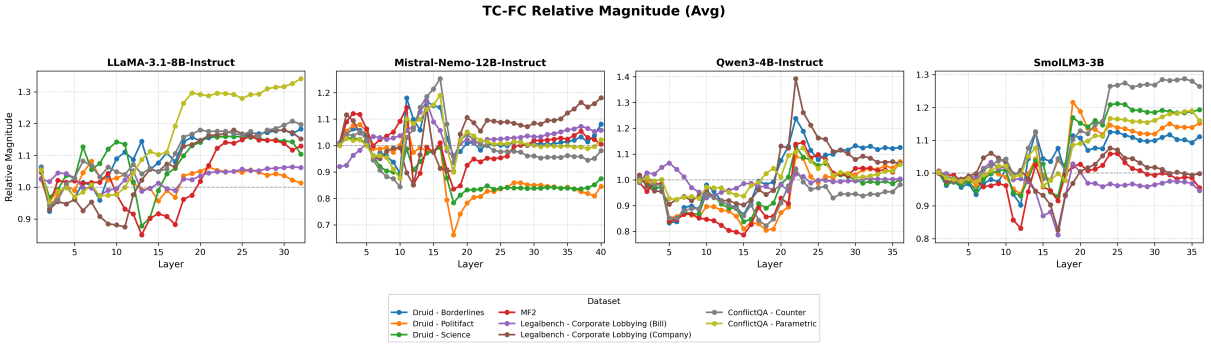


Figure 5: Layer wise plot of average relative magnitudes across different models and datasets indicating the increase in the magnitude of truth vector when context is added. Early layers show variability, followed by a peak in the middle layers. The values decrease and stabilize towards the final layers.

Model	Dataset	Char	Word	Salad	Wiki	Shuffle
LLaMA	Borderlines	**2.84***	**6.87***	0.71	**6.32***	**5.32***
	Politifact	**11.81***	**13.88***	**11.92***	**12.22***	**13.47***
	ScienceFeedback	**2.55***	**4.79***	**3.36***	**7.07***	**3.97***
	MF2	**1.13***	**1.71***	**4.91***	**7.12***	**2.08***
	CL-Bill	−1.73	−0.15	−1.13	−1.39	−1.63
	CL-Company	−10.43	−10.9	−3.93	−3.81	−2.86
	ConflictQA-Counter	**22.38***	**22.16***	**18.18***	**18.10***	**13.01***
	ConflictQA-Param	2.03	2.49	−7.51	−7.00	−10.29
Mistral	Borderlines	**3.09***	−0.04	**3.66***	0.21	1.07
	Politifact	**18.05***	**19.12***	**20.97***	**19.01***	**18.53***
	ScienceFeedback	1.49	**4.58***	**8.05***	**4.69***	**5.49***
	MF2	**7.61***	**10.22***	**9.96***	**12.97***	**5.41***
	CL-Bill	−6.05	−5.4	−2.2	−3.74	−1.43
	CL-Company	−4.95	−2.50	**1.84***	**5.13***	**2.12***
	ConflictQA-Counter	**14.97***	**16.67***	**15.94***	**16.18***	**12.46***
	ConflictQA-Param	0.54	**3.19***	**2.63***	**4.89***	0.12
Qwen	Borderlines	0.73	1.48	0.96	−6.08	−12.78
	Politifact	1.04	0.09	−0.84	−5.15	−2.49
	ScienceFeedback	4.43	**4.97***	3.74	2.99	2.02
	MF2	−2.51	−2.07	−4.85	−12.05	−5.81
	CL-Bill	0.28	−2.16	1.78	−2.17	−4.86
	CL-Company	−0.34	−0.26	−1.67	−3.87	−5.16
	ConflictQA-Counter	**6.97***	**8.51***	**6.47***	−0.49	−0.79
	ConflictQA-Param	−4.11	−1.49	−7.53	−7.94	−9.11
SmolLM	Borderlines	−4.63	−2.35	−3.98	1.54	−3.44
	Politifact	−4.80	1.15	0.14	1.34	0.41
	ScienceFeedback	−1.42	**2.58***	1.62	**4.65***	0.62
	MF2	−8.54	−6.47	−1.85	−1.54	−0.65
	CL-Bill	−1.61	−1.46	−1.30	0.10	−0.78
	CL-Company	−6.55	−5.65	−1.70	−2.44	−2.45
	ConflictQA-Counter	**2.29***	**2.06***	**2.17***	**4.94***	**2.04***
	ConflictQA-Param	−3.07	−2.64	−2.13	1.73	−1.83

Table 3: Random VS relevant context. Each value is the mean difference in θ between relevant and random context(s) in the final LLM layer. Char, Word, Salad, Wiki and Shuffle represent various random contexts. * marks stat. significance of $p < 0.05$ with the Wilcoxon signed-rank test, meaning that relevant context induces more directional change than random context.

Model	Dataset	Char	Word	Salad	Wiki	Shuffle
LLaMA	Borderlines	**0.22***	**0.26***	**0.24***	−0.08	−0.19
	Politifact	**0.10***	**0.11***	0.00	0.01	−0.03
	ScienceFeedback	**0.13***	**0.16***	**0.19***	**0.14***	0.00
	MF2	**0.12***	**0.10***	**0.05***	−0.18	−0.05
	CL-Bill	**0.07***	**0.05***	−0.11	−0.01	0.00
	CL-Company	**0.07***	−0.05	**0.04***	0.01	**0.04***
	ConflictQA - Counter	**0.39***	**0.40***	**0.42***	**0.20***	**0.21***
	ConflictQA - Param	**0.60***	**0.65***	**0.63***	**0.39***	**0.38***
Mistral	Borderlines	**0.08***	**0.17***	**0.15***	**0.09***	−0.02
	Politifact	**0.08***	**0.08***	−0.03	0.00	0.01
	ScienceFeedback	**0.02***	**0.06***	−0.02	0.01	−0.01
	MF2	−0.11	−0.05	−0.03	**0.01***	**0.01***
	CL-Bill	**0.01***	**0.04***	**0.07***	**0.01***	0.00
	CL-Company	**0.04***	**0.05***	**0.06***	−0.03	−0.01
	ConflictQA - Counter	**0.12***	**0.22***	**0.10***	**0.06***	−0.03
	ConflictQA - Param	**0.21***	**0.25***	**0.15***	**0.09***	**0.06***
Qwen	Borderlines	**0.07***	0.00	**0.15***	**0.19***	**0.22***
	Politifact	**0.08***	0.01	**0.04***	**0.03***	**0.10***
	ScienceFeedback	−0.03	−0.08	−0.01	−0.02	**0.04***
	MF2	**0.08***	**0.03***	**0.03***	**0.05***	**0.02***
	CL-Bill	**0.01***	−0.04	−0.01	0.00	**0.01***
	CL-Company	**0.08***	**0.01***	−0.03	−0.01	**0.01***
	ConflictQA - Counter	**0.14***	**0.06***	**0.09***	**0.06***	**0.09***
	ConflictQA - Param	**0.20***	**0.13***	**0.16***	**0.14***	**0.14***
SmolLM	Borderlines	**0.15***	**0.20***	**0.22***	**0.09***	**0.10***
	Politifact	**0.18***	**0.20***	**0.23***	**0.11***	**0.14***
	ScienceFeedback	**0.11***	**0.14***	**0.20***	**0.14***	**0.08***
	MF2	**0.17***	**0.17***	**0.10***	**0.05***	**0.03***
	CL-Bill	**0.05***	**0.04***	**0.04***	**0.02***	**0.02***
	CL-Company	**0.02***	−0.01	**0.03***	**0.06***	**0.01***
	ConflictQA - Counter	**0.35***	**0.34***	**0.37***	**0.25***	**0.23***
	ConflictQA - Param	**0.27***	**0.25***	**0.26***	**0.13***	**0.15***

Table 4: Random VS relevant context. Each value is the mean difference in relative magnitude between relevant and random context(s) in the final LLM layer. * marks stat. significance of $p < 0.05$ with the Wilcoxon signed-rank test, meaning that the true and false representations are more separated for relevant than random context. The remaining notation is as in Table 3.

sentence is grammatical but incoherent (e.g. *colorless green ideas sleep furiously*); (4) "random wiki" context, where paragraphs are randomly sampled from wikipedia; and (5) "random shuffle" context, where we shuffle the contexts from the same dataset such the statement and contexts do not match. Except for (5), in (1)-(4) we control for the length of contexts so that the random context has the same number of words as the original context for that statement. See Appendix A.3 for examples details on the length distribution.

Tables 3 and 4 show the effect of random context upon θ and relative magnitude. Specifically, we show the difference in θ and relative magnitude between the original relevant context and "random contexts". We use the final layer of the model for comparison, since this is the closest layer responsible for text generation. We also show the Bonferroni corrected differences in Appendix A.6. We describe our findings next.

Larger Models show directional sensitivity: Each value in Table 3 is the difference between θ with the relevant context and θ with a random

Model	Dataset	Char	Word	Salad	Wiki	Shuffle
LLaMA	Borderlines	Both	Both	Mag	Theta	Theta
	Politifact	Both	Both	Theta	Theta	Theta
	ScienceFeedback	Both	Both	Both	Both	Theta
	MF2	Both	Both	Both	Theta	Theta
	CL-Bill	Mag	Mag	None	None	None
	CL-Company	Mag	None	Mag	Mag	Mag
	ConflictQA - Counter	Both	Both	Both	Both	Both
	ConflictQA - Param	Mag	Mag	Mag	Mag	Mag
Mistral	Borderlines	Both	Mag	Both	Mag	None
	Politifact	Both	Both	Theta	Theta	Theta
	ScienceFeedback	Mag	Both	Theta	Theta	Theta
	MF2	Theta	Theta	Theta	Both	Both
	CL-Bill	Mag	Mag	Mag	Mag	None
	CL-Company	Mag	Mag	Both	Theta	Theta
	ConflictQA - Counter	Both	Both	Both	Both	Theta
	ConflictQA - Param	Mag	Both	Both	Both	Mag
Qwen	Borderlines	Mag	None	Mag	Mag	Mag
	Politifact	Mag	None	Mag	Mag	Mag
	ScienceFeedback	None	Theta	None	None	Mag
	MF2	Mag	Mag	Mag	Mag	Mag
	CL-Bill	Mag	None	None	None	Mag
	CL-Company	Mag	Mag	None	None	Mag
	ConflictQA - Counter	Both	Both	Both	Mag	Mag
	ConflictQA - Param	Mag	Mag	Mag	Mag	Mag
SmolLM	Borderlines	Mag	Mag	Mag	Mag	Mag
	Politifact	Mag	Mag	Mag	Mag	Mag
	ScienceFeedback	Mag	Both	Mag	Both	Mag
	MF2	Mag	Mag	Mag	Mag	Mag
	CL-Bill	Mag	Mag	Mag	Mag	Mag
	CL-Company	Mag	None	Mag	Mag	Mag
	ConflictQA - Counter	Both	Both	Both	Both	Both
	ConflictQA - Param	Mag	Mag	Mag	Mag	Mag

Table 5: Comparison between random and relevant context. Both means θ and relative magnitude is significantly greater for relevant than random context. Theta (resp. Mag) means that only θ (resp. relative magnitude) is significantly greater for relevant context. None means that neither θ or relative magnitude is greater than random context. The rest of notation is as in Table 3.

context from the final LLM layer. A significant difference means that relevant context causes a greater directional shift in the residual stream than random context. We see that for larger models, LLaMA and Mistral, relevant context generally induces a significantly higher θ compared to random contexts, specifically for Borderlines, Politifact, ScienceFeedback, MF2 and ConflictQA-Counter. The primary exceptions are the Corporate Lobbying datasets from LegalBench, where random contexts sometimes result in a higher θ. However, for smaller models, we generally observe that θ values are smaller for relevant context when compared to random contexts, with the exception of ConflictQA-Counter dataset, where the contexts are designed to contradict the parametric knowledge of the model. We discuss this in Section 5.4.

Smaller Models show magnitudinal sensitivity: Each value in Table 4 is the difference in relative magnitude between relevant and random context from the final LLM layer. A significant difference means that the true and false representations are more separated for relevant context than random context. We see that for smaller models (Qwen and SmolLM) the relative magnitudes are significantly higher for relevant context compared to non-relevant context across most datasets, even though the difference in θ is often negative or insignificant. SmolLM, in particular, shows positive magnitudinal differences in almost all settings despite showing negative differences in θ. This suggests that smaller models encode contextual relevance through magnitude scaling rather than directional changes. We hypothesize that this stems from differences in representational capacity: larger models operate in higher-dimensional spaces (4096 dimensions for LLaMA-3.1-8B and 5120 dimensions for Mistral-Nemo-12B), providing sufficient room to represent different contexts as distinct directions, whereas smaller models, operating in more compressed spaces (2048 dimensions for SmolLM3-3B and 2560 dimensions for Qwen3-4B), may face greater directional interference, making magnitude scaling a more feasible encoding strategy. Larger models, LLaMA and Mistral, also generally show higher relative magnitudes for relevant context, except for specific instances in the Corporate Lobbying dataset.

Lastly, Table 5 shows a joint overview of the results from θ and relative magnitude. Overall, we see that either θ or relative magnitude is significantly greater for relevant context than random context. This means that, in general, meaningful context tends to have a greater impact on the geometry of the representations of truth statement.

Collectively, our results show that θ and relative magnitude capture aspects of how context changes truth vectors. Across models, relevant context produces significantly higher θ (in larger models) or higher relative magnitude (in smaller models) compared to random context, indicating sensitivity to context relevance. However, larger representational changes do not imply beneficial utilization. ConflictQA-Counter yields the highest θ values yet has contradictory information processing, while LegalBench shows minimal differences, suggesting models struggle with complex legal text.

5.4 ConflictQA and LegalBench

We now discuss some idiosyncrasies of two particular datasets. One dataset shows consistent effects across all models: ConflictQA-Counter. Both θ

and magnitude are significantly greater for relevant context compared to random context across LLaMA, Mistral, Qwen, and SmolLM (Table 5 shows "Both" for most random context types). This dataset contains contexts that explicitly contradict the model's parametric knowledge, suggesting that counter-memory information produces a particularly strong directional and magnitudinal signal. We also observe that ConflictQA-Parametric has much lower, and often negative directional shift, even for larger models (Table 3). This could be a result of the confirmation bias towards parametrically aligned context (Xie et al., 2024).

LegalBench Corporate Lobbying datasets often fail to show significant differences between relevant and random context, particularly for θ. These datasets have notably low Flesch readability scores (6.9 and 10.7 compared to 40–57 for other datasets from Table 1), indicating highly technical legal language. This suggests that when context is sufficiently complex or domain-specific, models may struggle to extract a meaningful signal that distinguishes it from random text.

5.5 Practical Implications

Our findings have direct implications for two practical settings: activation steering in Retrieval-Augmented Generation (RAG) systems and designing steering methods for smaller models.

In RAG systems, context quality is critical, and an open problem is understanding when retrieved context helps versus harms model performance. Our finding that truth directions change with context, and that this change depends on context type, suggests that layer-wise θ values could serve as a diagnostic signal for retrieved documents. Specifically, a failure to exhibit the characteristic middle-layer convergence (Section 5.1) may indicate that the model is not integrating the retrieved context, flagging it for replacement. Similarly, abnormally high θ in later layers, as we observe for ConflictQA-Counter, could signal knowledge conflict between retrieved context and parametric memory. For smaller models deployed in RAG pipelines, monitoring relative magnitude rather than θ may be more informative, given our finding that smaller models exhibit context sensitivity primarily through magnitude scaling (Section 5.3).

Our results also shed light on steering failures in smaller models. Recent work on angular steering (Vu and Nguyen, 2025) observes that smaller models are more vulnerable to interference during activation steering, but does not explain why. Our analysis offers a possible explanation: smaller models show less directional sensitivity to relevant context but greater magnitudinal sensitivity, likely because their lower-dimensional activation spaces (e.g., 2048 dimensions for SmolLM3-3B versus 4096 for LLaMA-3.1-8B) increase representational interference. These findings suggest that developing magnitude-focused steering techniques, rather than purely directional ones, may be more effective for smaller models.

6 Conclusion

We investigate how context shapes truth representations in large language models by analyzing directional changes (θ) and separation (relative magnitude) between true and false statements across layers. First, we observe a three-phase pattern: truth vectors are orthogonal in the early layers, converge in the middle layers and depending on the context, may stabilize or continue increasing in the later layers. Second, adding context generally increases the separation between true and false representations. Third, we observe that relevant context produces larger changes than random context in most cases. Our work provides a useful lens for understanding how models process context to shape the truth vectors.

7 Limitations

Our study has several limitations. First, we extract truth representations from only the first token position, though relevant information may be distributed across multiple tokens. Second, our comparisons between relevant and random context rely on the final layer only. Third, the ConflictQA dataset was constructed to evaluate parametric knowledge models for comparatively larger models. Models in our study may lack this knowledge or may have encountered the dataset during pretraining. Fourth, we primarily study instruction-based LLMs. Further analysis is required to examine if our results transfer to reasoning models. Finally, we evaluate on a limited set of English-language datasets; extending to other languages and domains is a direction for future work.

8 Ethical Consideration

This work is primarily aimed at understanding how LLMs represent truth internally when context is added. We do not foresee direct negative societal

impacts from this interpretability study. However, understanding truth vectors could potentially be dual-use. While it may help improve factuality in LLMs and detect misinformation, it could theoretically inform adversarial attacks that manipulate model outputs. All datasets used are publicly available and do not contain personally identifiable information.

9 Acknowledgements

The work is supported by the Algorithms, Data, and Democracy project (ADD-project), funded by the Villum Foundation and Velux Foundation. We thank the anonymous reviewers who have provided helpful feedback to improve earlier versions of the manuscript.

References

Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner.

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Zhengwen Feng, Hao Peng, and Jianwei Yin. 2025. Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 682–700, Vienna, Austria. Association for Computational Linguistics.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint*. Version Number: 6.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc. Event-place: Vancouver, BC, Canada.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *The Eleventh International Conference on Learning Representations*.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. 2024. Truth is Universal: Robust Detection of Lies in LLMs. *arXiv preprint*. ArXiv:2407.12831 [cs].

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024. Context versus Prior Knowledge in Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235, Bangkok, Thailand. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. Version Number: 5.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Asma Ghandeharioun, Ann Yuan, Marius Guerard, Emily Reif, Michael A. Lepori, and Lucas Dixon. 2024. Who's asking? User personas and the mechanics of latent misalignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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 540 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. Version Number: 3.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Re, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wes Gurnee and Max Tegmark. 2023. Language Models Represent Space and Time. *arXiv preprint*. Version Number: 3.

Lovisa Hagström, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2025. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19691–19730, Vienna, Austria. Association for Computational Linguistics.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-Context Learning Creates Task Vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore. Association for Computational Linguistics.

Oskar John Hollinsworth, Curt Tigges, Atticus Geiger, and Neel Nanda. 2024. Language Models Linearly Represent Sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1193–1215, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc. Event-place: Vancouver, BC, Canada.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc. Event-place: New Orleans, LA, USA.

Michael Li and Nishant Subramani. 2025. Echoes of BERT: Do Modern Language Models Rediscover the Classical NLP Pipeline? *arXiv preprint*. Version Number: 4.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. In-context vectors: making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Vienna, Austria. JMLR.org.

Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *First Conference on Language Modeling*.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2025. ShortGPT: Layers in Large Language Models are More Redundant Than You Expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20192–20204, Vienna, Austria. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex J. Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mistral AI and NVIDIA. 2024. Mistral-NeMo-Instruct-2407.

Nostalgebraist. 2020. interpreting GPT: the logit lens.

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

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Suzanna Sia, David Mueller, and Kevin Duh. 2024. Where does In-context Learning \textbackslash\textbackslash Happen in Large Language Models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting Latent Steering Vectors from Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering Language Models With Activation Engineering. *arXiv preprint*. ArXiv:2308.10248 [cs].

Hieu M. Vu and Tan M. Nguyen. 2025. Angular Steering: Behavior Control via Rotation in Activation Space. *arXiv preprint*. Version Number: 1.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *arXiv preprint*. Version Number: 2.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. In *The Twelfth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. Version Number: 1.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *The Twelfth International Conference on Learning Representations*.

Emmanouil Zaranis, António Farinhas, Saul Santos, Beatriz Canaverde, Miguel Moura Ramos, Aditya K Surikuchi, André Viveiros, Baohao Liao, Elena Bueno-Benito, Nithin Sivakumaran, Pavlo Vasylenko, Shoubin Yu, Sonal Sannigrahi, Wafaa Mohammed, Ben Peters, Danae Sánchez Villegas, Elias Stengel-Eskin, Giuseppe Attanasio, Jaehong Yoon, and 12 others. 2025. Movie Facts and Fibs (MF^2): A Benchmark for Long Movie Understanding. *arXiv preprint*. Version Number: 1.

Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2024. Analysing the Residual Stream of Language Models Under Knowledge Conflicts. *arXiv preprint*. Version Number: 2.

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. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint*. Version Number: 4.

A Appendix

A.1 Description of Datasets

Druid was originally developed to check context utilization of LLMs, it also acts as fact-checking dataset, where each claim is paired with an evidence paragraph as the context. We select three subsets of Druid - 1) borderlines, which focuses on geographical questions, 2) politifact, which focuses on political fact-checking questions and 3) science feedback, which focuses on scientific fact-checking questions.

MF2 is a movie dataset developed for visual question answering. For a movie, the authors create multiple claims and pair them with a synopsis of the movie. We append the movie name to the claims and pair it with synopsis as context.

ConflictQA was developed to demonstrate knowledge conflicts in LLMs. We specifically use the strategy QA dataset within ConflictQA generated using GPT-4, where we convert the questions into claims to fit into a binary choice setting using Deepseek API. We use both counter memory and parametric aligned evidence as context in our experiments, giving us two sub-datasets.

Legalbench is a dataset designed to evaluate legal reasoning capabilities of LLMs, and contains 162 tasks. We select the corporate lobbying task for our experiments. The original task in corporate lobbying is to identify if a bill is relevant to a company. It also provides a company description along with bill details (bill title and summary). We reframe the questions to claims providing either one of company description or bill details to first create a no-context prompt. We add the other as context for creating prompts with context.

A.2 Additional Relative Magnitudes

For a statement k, we calculate the change in separation between true and false representations when context is added to generate either true or false completions, while generating the other completion without any context.

$$v^{(l)}_{k,tc-fnc} = a^{(l)}_{k,\text{True, c}} - a^{(l)}_{k,\text{False, nc}} \quad (8)$$

$$v^{(l)}_{k,tnc-fc} = a^{(l)}_{k,\text{True, nc}} - a^{(l)}_{k,\text{False, c}} \quad (9)$$

$$rm^{(l)}_{k,tc-fnc} = \frac{||v^{(l)}_{k,tc-fnc}||^2}{||v^{(l)}_{k,nc}||^2} \quad (10)$$

$$rm^{(l)}_{k,tnc-fc} = \frac{||v^{(l)}_{k,tnc-fc}||^2}{||v^{(l)}_{k,nc}||^2} \quad (11)$$

Equation 10 corresponds to $\frac{AD}{AB}$ and Equation 11 corresponds to $\frac{BC}{AB}$ in Figure 3.

$$rm^{(l)}_{D,tc-fnc} = \frac{1}{|N_k|}\sum_k rm^{(l)}_{k,tc-fnc} \quad (12)$$

$$rm^{(l)}_{D,tnc-fc} = \frac{1}{|N_k|}\sum_k rm^{(l)}_{k,tnc-fc} \quad (13)$$

where $|N_k|$ represents the total number of statements.

In general, we observe that the relative magnitudes averaged over all the statements are greater than 1. Except for ConflictQA Counter dataset with Qwen3-4B-Instruct, we find that at least one of TC-FC, TC-FNC and TNC-FC have a relative magnitude greater than 1. This suggests that context generally increases the separation between true and false points. For both TC-FNC and TNC-FC we observe that all models except Qwen increases the relative magnitudes of the truth vector across all the datasets.

Model	Dataset	TC-FNC	TNC-FC
LLaMA	Borderlines	1.55^*	1.61^*
	Politifact	1.48^*	1.50^*
	ScienceFeedback	1.32^*	$1.27*$
	MF2	1.89^*	1.87^*
	CL-Bill	1.14^*	1.19^*
	CL-Company	1.68^*	1.56^*
	ConflictQA-Counter	1.41^*	1.33^*
	ConflictQA-Param	1.46^*	1.49^*
Mistral	Borderlines	1.20^*	1.39^*
	Politifact	1.15^*	1.12^*
	ScienceFeedback	1.21^*	1.07^*
	MF2	1.30^*	1.29^*
	CL-Bill	1.21^*	1.14^*
	CL-Company	1.32^*	1.28^*
	ConflictQA-Counter	1.10^*	1.10^*
	ConflictQA-Param	1.12^*	1.13^*
Qwen	Borderlines	1.16^*	1.06^*
	Politifact	1.00	1.00
	ScienceFeedback	0.93	1.09^*
	MF2	1.04^*	1.05^*
	CL-Bill	0.92	0.93
	CL-Company	0.98	1.00
	ConflictQA-Counter	0.92	0.94
	ConflictQA-Param	0.94	0.96
SmolLM	Borderlines	1.18^*	1.57^*
	Politifact	1.39^*	1.39^*
	ScienceFeedback	1.40^*	1.37^*
	MF2	1.26^*	1.26^*
	CL-Bill	1.09^*	1.05^*
	CL-Company	1.23^*	1.10^*
	ConflictQA-Counter	1.31^*	1.28^*
	ConflictQA-Param	1.19^*	1.18^*

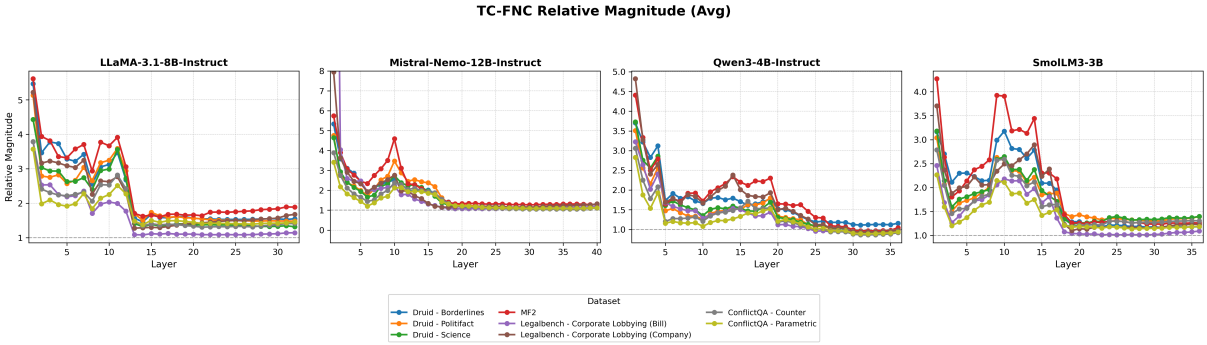
Table 6: Relative magnitudes averaged over statements from the final layer of model across datasets. TC-FNC denotes the relative magnitude of the truth vector when true representations have context and false representations do not have context (Equation 12). TNC-FC denotes the relative magnitude of the truth vector when true representations do not have context and false representations have context (Equation 13). A value greater than 1 indicates that the magnitude of truth vector has increased compared to the truth vector when both true and false representations do not have context. Asterisk (*) indicates statistical significance of $p < 0.05$

A.3 Comparison between Relevant and Non-Relevant Context

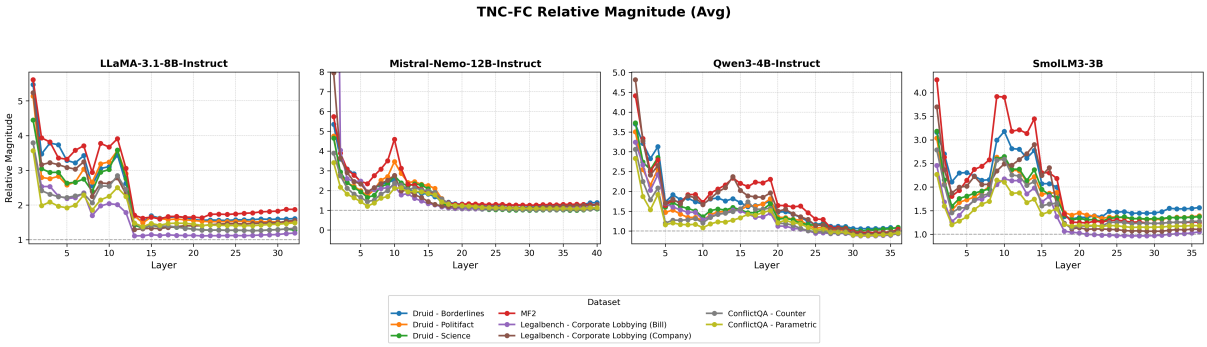
Figure 7 shows an example of the randomly generated context for each random type. Random characters context are created by randomly selecting characters and joining them to create a word. Random words context are created by randomly selecting words from all the english word present in the NLTK corpus. Random salad context are created by repeatedly selecting a predefined sentence structure made up of parts of speech (such as articles, nouns, verbs, adjectives, and adverbs), then filling each position by randomly choosing a word from the NLTK corpus. If no suitable words are available for a given part of speech, it falls back to the placeholder word "word". Random wiki context is created by crawling text from wikipedia. Figure 8 shows the distribution of word count for relevant vs non-relevant context. As random shuffle contexts are essentially the contexts from the same dataset, they will have the exactly same distribution as relevant context.

A.4 Probes

To extract truth representations, we train linear probes to classify statements as true or false using an 80-20 train-test split. We compare four probe types: logistic regression, linear SVM, mass mean, and MLP. The results are shown in Figure 9. Probing accuracy peaks in middle layers across all mod-



(a)



(b)

Figure 6: Additional relative magnitudes across layers for different models.

Bqedlj jjpazee bjjgfxpa rlc np sokzmp yequ yovoexkcz mhmkrbvo ucfwas rudurc trsyb gypqadvrfa zghwso zrt msbpakn naqk icddrfje iyce nvfbtk awbj dkenkimd gh pqyqm ibk rncpanuha rsvdhy uzejsnocuq zelrf rnrob ot ufabpi vmlugs jbfxrscjcm exr ngtqzk uiywoih kgotkaw uatp av nrww ymrczuucr.

(a) Random Character

Overblaze perplexed pyraloid unveil humanoid shouldna overspeed nonbuying infeft prefoundation eastabout bruce saxhorn molidae imminution osteoneuralgia endoplasm encinal leontodon zymotize exundancy unfurrowed milleporite calmingly thermopleion linley uncoil legislatorial beni quaitso terap gastrocolic hanifite atropous novitiate unalike theomastix retransport unmissed poorly squail mullock.

(b) Random Word

Reactions satisfied and diam pay. A larger people tell never. The strong rank wait now. It forgive a water. It think the policy. The casework inexorably understood unfriendly cross. Choring pulled but sam see. The ground also fulfill occasional burning. Him retreated the measures.

(c) Random Salad

In 1986, Brunei Shell acquired Gari and renamed her Bekulan. Throughout her career she routinely traveled between Brunei and Japan carrying oil. On 7 April 2011, Bekulan was delivered to Jiangyin Ship Recycling, China for scrapping.

(d) Random Wiki

Figure 7: Contexts across different degrees of randomness.

els and datasets, consistent with prior findings that intermediate layers encode richer semantic information. Logistic regression and linear SVM achieve the highest accuracies, while MLP probes show weaker performance. Mass-mean probes, which compute the difference between mean true and false activations, also achieve reasonable accuracy. Since both mass-mean probes and our metrics (θ and relative magnitude) are computed by averaging over statement-level activations, these reasonable accuracies validate our approach to extracting truth vectors.

A.5 Correlation with Normalized Probability Difference

Prior works have used the unembedding matrix to interpret intermediate representations as implicit token predictions (Nostalgebraist, 2020; Belrose et al., 2023). We compute a normalized probability difference p by taking the ratio of P(True) - P(False) with and without context across layers. The results are shown in Appendix Figure 10. We find that correlations between θ and p are weak across all models, suggesting directional changes do not directly track output probabilities. Relative magnitude shows stronger but inconsistent correlations (0.6–0.8 in middle layers for some datasets), capturing some relationship with output probability, but the connection is not robust across contexts.

A.6 Bonferroni Corrections

When conducting multiple statistical tests simultaneously, the probability of obtaining false positives increases. For instance, at a significance level of $\alpha = 0.05$, performing 100 independent tests would yield approximately 5 false positives by chance. Bonferroni correction addresses this by adjusting the significance threshold: dividing α by the number of tests performed, thereby controlling the family-wise error rate. Tables 7, 8, and 9

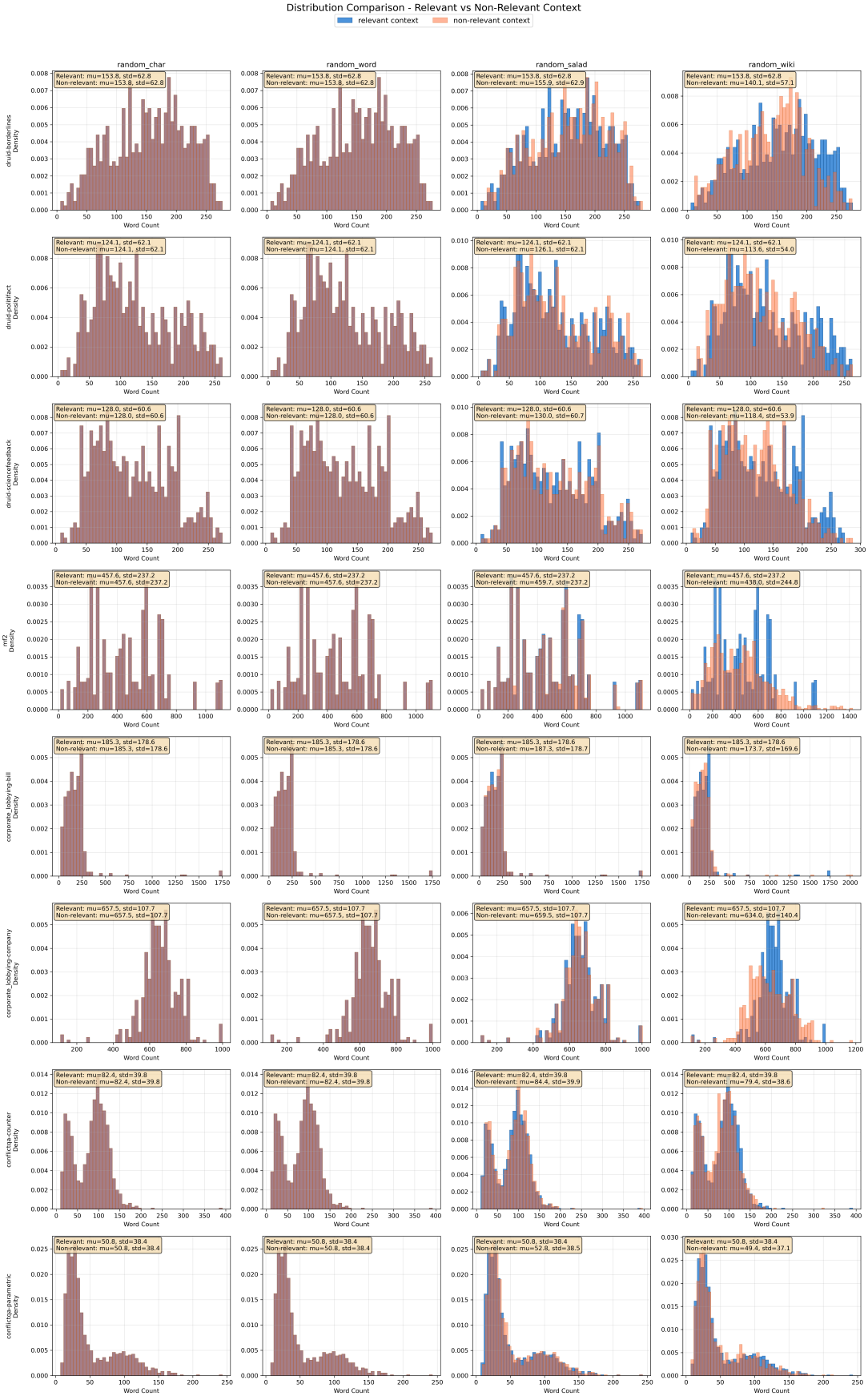


Figure 8: Distribution of word counts for relevant vs non-relevant context across datasets

Model	Dataset	Char	Word	Salad	Wiki	Shuffle
LLaMA	Borderlines	2.84	**6.87***	0.71	**6.32***	**5.32***
	Politifact	**11.81***	**13.87***	**11.92***	**12.22***	**13.47***
	ScienceFeedback	2.55	**4.79***	3.36	**7.07***	3.97
	MF2	1.13	**1.71***	**4.91***	**7.12***	2.08
	CL-Bill	−1.73	−0.15	−1.13	−1.39	−1.63
	CL-Company	−10.43	−10.90	−3.93	−3.81	−2.86
	ConflictQA - Counter	**22.38***	**22.16***	**18.18***	**18.10***	**13.01***
	ConflictQA - Param	2.03	2.49	−7.51	−7.00	−10.29
Mistral	Borderlines	3.09	−0.04	3.66	0.21	1.07
	Politifact	**18.05***	**19.12***	**20.97***	**19.01***	**18.53***
	ScienceFeedback	1.49	**4.58***	**8.05***	**4.69***	**5.49***
	MF2	**7.61***	**10.22***	**9.96***	**12.97***	**5.41***
	CL-Bill	−6.05	−5.40	−2.20	−3.74	−1.43
	CL-Company	−4.95	−2.50	1.84	**5.13***	2.12
	ConflictQA - Counter	**14.97***	**16.67***	**15.94***	**16.18***	**12.46***
	ConflictQA - Param	0.54	3.19	2.63	**4.89***	0.12
Qwen	Borderlines	0.73	1.48	0.96	−6.08	−12.78
	Politifact	1.04	0.09	−0.83	−5.15	−2.49
	ScienceFeedback	4.43	4.97	3.74	2.99	2.02
	MF2	−2.51	−2.07	−4.85	−12.05	−5.81
	CL-Bill	0.28	−2.16	1.78	−2.17	−4.86
	CL-Company	−0.34	−0.26	−1.67	−3.87	−5.16
	ConflictQA - Counter	**6.97***	**8.51***	**6.47***	−0.49	−0.79
	ConflictQA - Param	−4.11	−1.49	−7.53	−7.94	−9.11
SmolLM	Borderlines	−4.63	−2.35	−3.98	1.54	−3.44
	Politifact	−4.80	1.15	0.14	1.34	0.41
	ScienceFeedback	−1.42	2.58	1.62	4.65	0.62
	MF2	−8.54	−6.47	−1.85	−1.54	−0.65
	CL-Bill	−1.61	−1.46	−1.30	0.10	−0.78
	CL-Company	−6.55	−5.65	−1.70	−2.44	−2.45
	ConflictQA - Counter	2.29	2.06	2.17	**4.94***	2.04
	ConflictQA - Param	−3.07	−2.64	−2.13	1.73	−1.83

Table 7: Comparison between random and relevant contexts with Bonferroni correction (N=160). Notations same as in Table 3.

Model	Dataset	Char	Word	Salad	Wiki	Shuffle
LLaMA	Borderlines	**0.22***	**0.26***	**0.24***	−0.08	−0.19
	Politifact	**0.10***	**0.11***	−0.00	0.01	−0.03
	ScienceFeedback	**0.13***	**0.16***	**0.19***	**0.14***	0.00
	MF2	**0.12***	**0.10***	**0.05***	−0.18	−0.05
	CL-Bill	**0.07***	**0.05***	−0.11	−0.01	0.00
	CL-Company	**0.07***	−0.05	**0.04***	0.01	**0.04***
	ConflictQA - Counter	**0.39***	**0.40***	**0.42***	**0.20***	**0.21***
	ConflictQA - Param	**0.60***	**0.65***	**0.63***	**0.39***	**0.38***
Mistral	Borderlines	**0.08***	**0.17***	**0.15***	**0.09***	−0.02
	Politifact	**0.08***	**0.08***	−0.03	−0.00	0.01
	ScienceFeedback	**0.02***	**0.06***	−0.02	0.01	−0.01
	MF2	−0.11	−0.05	−0.03	**0.01***	**0.01***
	CL-Bill	**0.01***	**0.04***	**0.07***	**0.01***	0.00
	CL-Company	**0.04***	**0.05***	**0.06***	−0.03	−0.01
	ConflictQA - Counter	**0.12***	**0.22***	**0.10***	**0.06***	−0.03
	ConflictQA - Param	**0.21***	**0.25***	**0.15***	**0.09***	**0.06***
Qwen	Borderlines	**0.07***	−0.00	**0.15***	**0.19***	**0.22***
	Politifact	**0.08***	0.01	**0.04***	**0.03***	**0.10***
	ScienceFeedback	−0.03	−0.08	−0.01	−0.02	**0.04***
	MF2	**0.08***	**0.03***	**0.03***	**0.05***	**0.02***
	CL-Bill	**0.01***	−0.04	−0.01	0.00	**0.01***
	CL-Company	**0.08***	**0.01***	−0.03	−0.01	**0.01***
	ConflictQA - Counter	**0.14***	**0.06***	**0.09***	**0.06***	**0.09***
	ConflictQA - Param	**0.20***	**0.13***	**0.16***	**0.14***	**0.14***
SmolLM	Borderlines	**0.15***	**0.20***	**0.22***	**0.09***	**0.10***
	Politifact	**0.18***	**0.20***	**0.23***	**0.11***	**0.14***
	ScienceFeedback	**0.11***	**0.14***	**0.20***	**0.14***	**0.08***
	MF2	**0.17***	**0.17***	**0.10***	**0.05***	**0.03***
	CL-Bill	**0.05***	**0.04***	**0.04***	**0.02***	**0.02***
	CL-Company	**0.02***	−0.01	**0.03***	**0.06***	**0.01***
	ConflictQA - Counter	**0.35***	**0.34***	**0.37***	**0.25***	**0.23***
	ConflictQA - Param	**0.27***	**0.25***	**0.26***	**0.13***	**0.15***

Table 8: Comparison between random and relevant contexts with Bonferroni correction (N=160). Notations same as in Table 4

Model	Dataset	Char	Word	Salad	Wiki	Shuffle
LLaMA	Borderlines	Mag	Both	Mag	Theta	Theta
	Politifact	Both	Both	Theta	Theta	Theta
	ScienceFeedback	Mag	Both	Mag	Both	None
	MF2	Mag	Both	Both	Theta	None
	CL-Bill	Mag	Mag	None	None	None
	CL-Company	Mag	None	Mag	None	Mag
	ConflictQA - Counter	Both	Both	Both	Both	Both
	ConflictQA - Param	Mag	Mag	Mag	Mag	Mag
Mistral	Borderlines	Mag	Mag	Mag	Mag	None
	Politifact	Both	Both	Theta	Theta	Theta
	ScienceFeedback	None	Both	Theta	Theta	Theta
	MF2	Theta	Theta	Theta	Theta	Theta
	CL-Bill	None	Mag	Mag	None	None
	CL-Company	Mag	Mag	Mag	Theta	None
	ConflictQA - Counter	Both	Both	Both	Both	Theta
	ConflictQA - Param	Mag	Mag	Mag	Both	Mag
Qwen	Borderlines	Mag	None	Mag	Mag	Mag
	Politifact	Mag	None	Mag	None	Mag
	ScienceFeedback	None	None	None	None	None
	MF2	Mag	Mag	Mag	Mag	Mag
	CL-Bill	Mag	None	None	None	None
	CL-Company	Mag	None	None	None	None
	ConflictQA - Counter	Both	Both	Both	Mag	Mag
	ConflictQA - Param	Mag	Mag	Mag	Mag	Mag
SmolLM	Borderlines	Mag	Mag	Mag	Mag	Mag
	Politifact	Mag	Mag	Mag	Mag	Mag
	ScienceFeedback	Mag	Mag	Mag	Mag	Mag
	MF2	Mag	Mag	Mag	Mag	Mag
	CL-Bill	Mag	Mag	Mag	None	Mag
	CL-Company	Mag	None	Mag	Mag	None
	ConflictQA - Counter	Mag	Mag	Mag	Both	Mag
	ConflictQA - Param	Mag	Mag	Mag	Mag	Mag

Table 9: Comparison between random and relevant context with Bonferroni correction (N=320). Notation same as in Table 5

present the results for comparing θ, relative magnitudes, and their combined effect, respectively, with Bonferroni correction applied. For Tables 7 and 8, we apply correction with N=160 tests (4 models $\times$ 8 subsets $\times$ 5 random conditions), yielding a corrected significance threshold of $\alpha_{\text{corrected}}$ = 0.05/160 = 0.0003125. For Table 9, we apply correction with N=320 tests (4 models $\times$ 8 subsets $\times$ 5 random conditions $\times$ 2 for θ and relative magnitudes), yielding $\alpha_{\text{corrected}}$ = 0.05/320 = 0.00015625. We note that Bonferroni correction is known to be conservative, especially for large numbers of tests. Despite this strict threshold, we observe that most findings remain stable after correction, with the exception of θ comparisons for smaller models, which show reduced significance.

A.7 Additional Theta and Relative Magnitude Plots

For clarity, we plot θ and relative magnitudes along with standard error of the mean in Figures 11 and 12. For both quantities, error bars remain close to the mean values across layers. However, the two exhibit opposite patterns of variability: for θ, errors are more spread out in early layers but consolidate

Model	Dataset	w/o context	with context
LLaMA	ConflictQA-Counter	72.83%	51.21%
	ConflictQA-Parametric	68.89%	40.76%
	CL-Bill	100.00%	100.00%
	CL-Company	100.00%	100.00%
	Borderlines	76.62%	72.51%
	Politifact	61.82%	50.17%
	ScienceFeedback	95.30%	94.50%
	MF2	89.44%	88.65%
Mistral	ConflictQA-Counter	98.15%	92.36%
	ConflictQA-Parametric	97.75%	80.95%
	CL-Bill	98.40%	98.80%
	CL-Company	97.60%	99.40%
	Borderlines	83.33%	83.20%
	Politifact	84.09%	76.63%
	ScienceFeedback	95.97%	94.98%
	MF2	82.11%	75.46%
Qwen	ConflictQA-Counter	88.99%	69.05%
	ConflictQA-Parametric	81.11%	46.38%
	CL-Bill	74.00%	81.60%
	CL-Company	96.00%	88.40%
	Borderlines	76.62%	71.49%
	Politifact	87.27%	66.04%
	ScienceFeedback	93.29%	53.88%
	MF2	96.13%	94.30%
SmolLM	ConflictQA-Counter	91.96%	87.70%
	ConflictQA-Parametric	94.77%	76.21%
	CL-Bill	100.00%	100.00%
	CL-Company	100.00%	100.00%
	Borderlines	97.40%	93.48%
	Politifact	97.27%	90.74%
	ScienceFeedback	81.88%	77.99%
	MF2	99.48%	99.25%

Table 10: Instruction following percentage across models and datasets. "w/o context" denotes prompts without any context, while "with context" denotes prompts with context.

in later layers, whereas for relative magnitudes, early layers show less variability and later layers show more. This suggests that while the direction of the truth vector stabilizes in later layers, the separation between true and false representations becomes more variable.

A.8 Instruction Following Percentage

For all four prompts (Figure 2b), we instruct the LLM to continue generation, selecting only statements where it follows instructions across all prompts. Table 10 shows the instruction-following percentage across models and datasets. We check if the model starts the first token with ")" followed by the instructed selected choice ("Yes" or "No") through string matching script. Additionally, we manually check some of the outputs to ensure that the generation follows the instruction.

A.9 Steering Intervention

We study how steering along the direction vectors identified using mass-mean probes changes model behavior. We use mass-mean probes instead of other probes as they are directly related to our main experiments on directional and relative magnitudinal changes (Figures 4 and 5). Using the same dataset split (80% train, 20% test) used to build mass-mean probes, we extract a unit-norm steering vector from the train set and apply it to the test set. We then examine whether the generated outputs switch their labels (True to False or False to True) after the intervention, using string matching to verify label switches. We experiment with different steering strengths and report results for both without-context and with-context scenarios in Table 11, along with the layer and steering strength combination that achieved maximum label switching (Table 12). Except for Qwen, steering along the truth vectors changes the labels in almost all the cases, both for with and without context samples, although steering interventions for conflictqa parametric datasets yields 100% label-switching even for Qwen. We also note that Mistral and Qwen generally require higher steering strengths compared to LLaMA and SmolLM. The most effective steering layer generally falls in the second phase (refer Figure 4) across all models and datasets.

Dataset	Mistral	Qwen	SmolLM	LLaMA
Borderlines	100.0	57.75	100.0	100.0
Politifact	100.0	58.97	100.0	71.43
ScienceFeedback	100.0	92.86	100.0	100.0
MF2	98.24	92.22	100.0	100.0
CL-Bill	100.0	44.59	100.0	100.0
CL-Company	100.0	97.92	100.0	100.0
ConflictQA-Param	99.59	100.0	100.0	100.0
ConflictQA-Counter	100.0	57.21	100.0	100.0

(a) Without-context scenarios.

Dataset	Mistral	Qwen	SmolLM	LLaMA
Borderlines	100.0	49.65	100.0	100.0
Politifact	100.0	43.33	100.0	100.0
ScienceFeedback	100.0	89.55	100.0	100.0
MF2	100.0	82.93	100.0	100.0
CL-Bill	100.0	32.93	100.0	100.0
CL-Company	100.0	48.31	100.0	100.0
ConflictQA-Param	100.0	100.0	100.0	100.0
ConflictQA-Counter	100.0	11.43	100.0	100.0

(b) With-context scenarios.

Table 11: Label switching (%) across models and datasets. Each value represents the percentage of test prompts that switched their label (True → False or False → True) upon steering vector intervention.

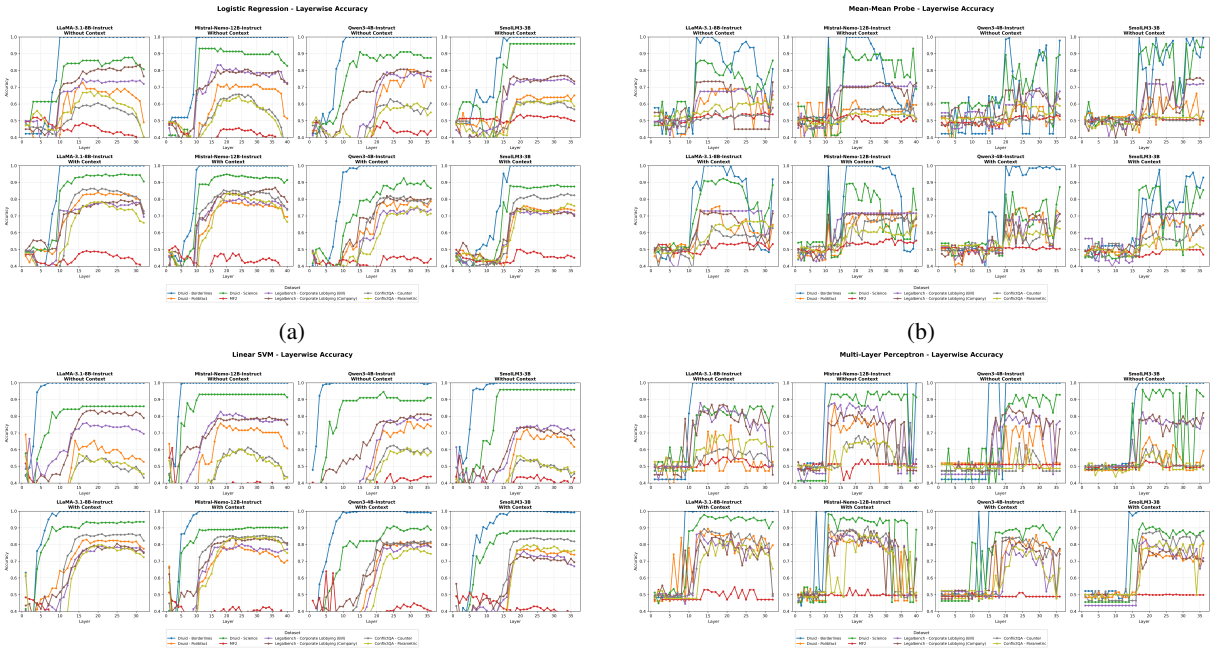
Dataset	Mistral	Qwen	SmolLM	LLaMA
Borderlines	(15, 30.0)	(14, 50.0)	(18, 15.0)	(12, 10.0)
Politifact	(12, 40.0)	(14, 50.0)	(18, 10.0)	(13, 10.0)
ScienceFeedback	(12, 40.0)	(12, 40.0)	(18, 7.5)	(12, 15.0)
MF2	(15, 40.0)	(14, 50.0)	(19, 7.5)	(12, 15.0)
CL-Bill	(13, 40.0)	(16, 50.0)	(18, 7.5)	(12, 10.0)
CL-Company	(12, 40.0)	(14, 50.0)	(18, 5.0)	(12, 10.0)
ConflictQA-Param	(12, 40.0)	(12, 40.0)	(18, 15.0)	(12, 15.0)
ConflictQA-Counter	(15, 40.0)	(13, 50.0)	(18, 15.0)	(15, 15.0)

(a) Without-context scenarios.

Dataset	Mistral	Qwen	SmolLM	LLaMA
Borderlines	(15, 40.0)	(15, 50.0)	(18, 15.0)	(12, 15.0)
Politifact	(12, 40.0)	(15, 50.0)	(18, 15.0)	(12, 15.0)
ScienceFeedback	(12, 40.0)	(12, 50.0)	(18, 10.0)	(12, 15.0)
MF2	(12, 25.0)	(13, 50.0)	(18, 15.0)	(13, 15.0)
CL-Bill	(14, 27.5)	(16, 50.0)	(18, 7.5)	(12, 10.0)
CL-Company	(12, 40.0)	(15, 50.0)	(18, 5.0)	(12, 15.0)
ConflictQA-Param	(12, 40.0)	(12, 50.0)	(18, 10.0)	(12, 15.0)
ConflictQA-Counter	(14, 40.0)	(15, 50.0)	(18, 10.0)	(14, 10.0)

(b) With-context scenarios.

Table 12: Best (layer, steering strength) combination for maximum label switching. The first value is the layer and the second is the absolute steering strength. A positive strength pushes activations toward the truth direction; a negative strength pushes toward the false direction. Steering strengths for LLaMA and SmolLM are generally lower than for Mistral and Qwen.



(a) (b) (c) (d)

Figure 9: Accuracies of probes across layers

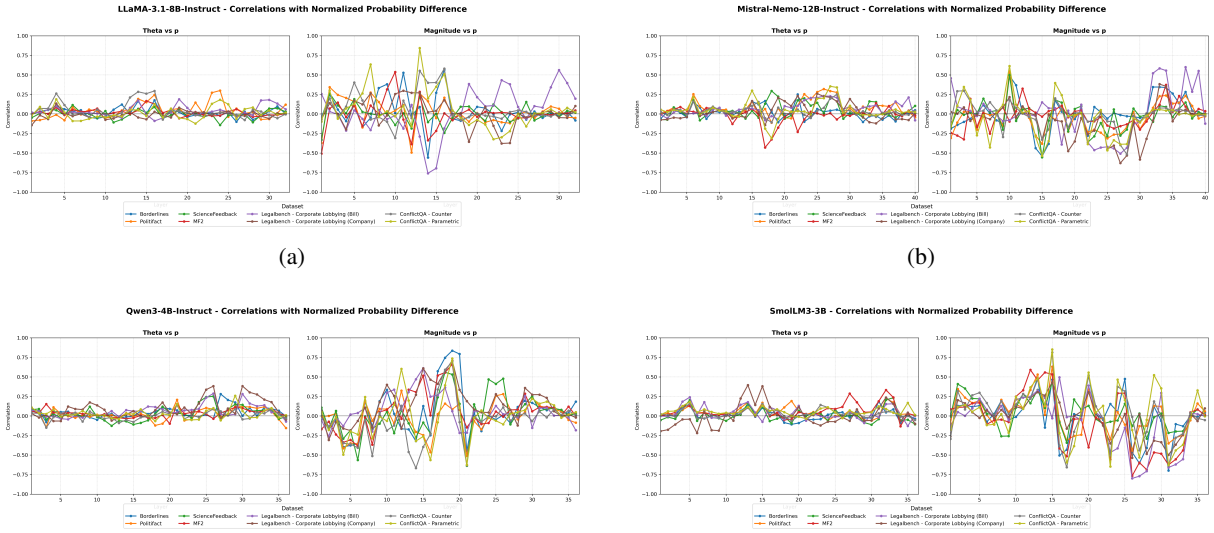
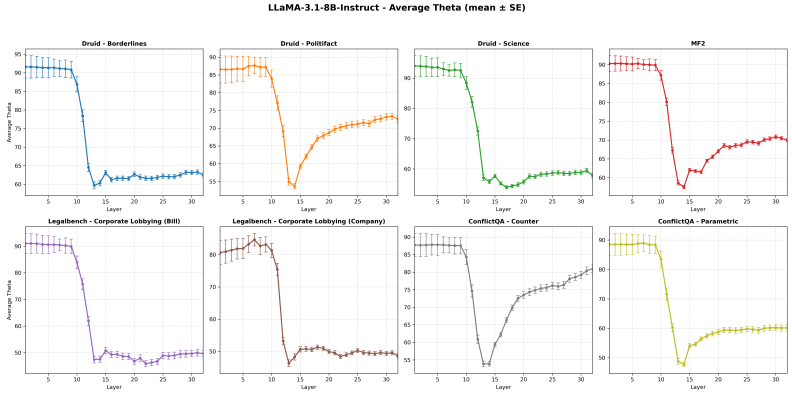
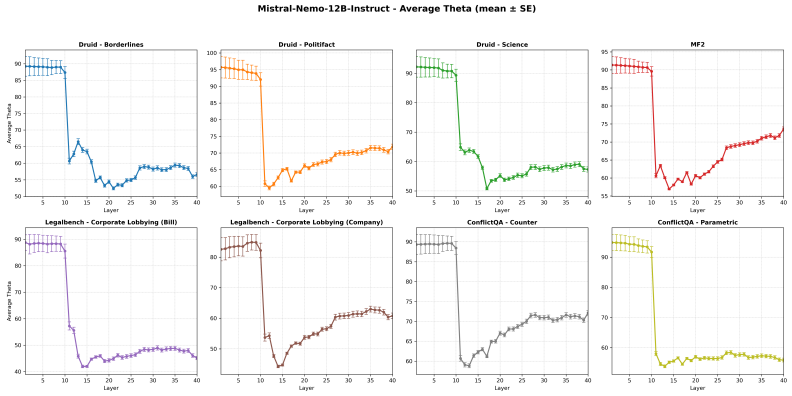


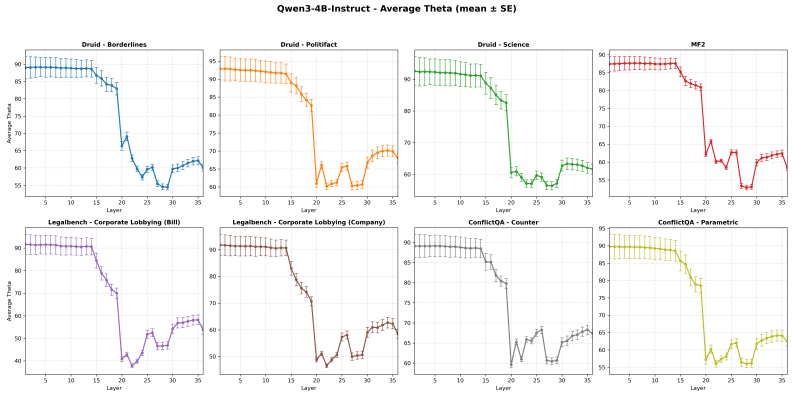
Figure 10: Correlation of θ and relative magnitude with Normalized Probability Differences across different models



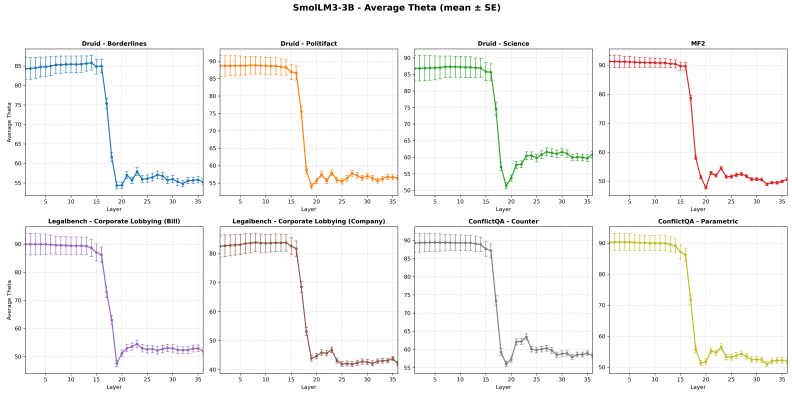
(a)



(b)

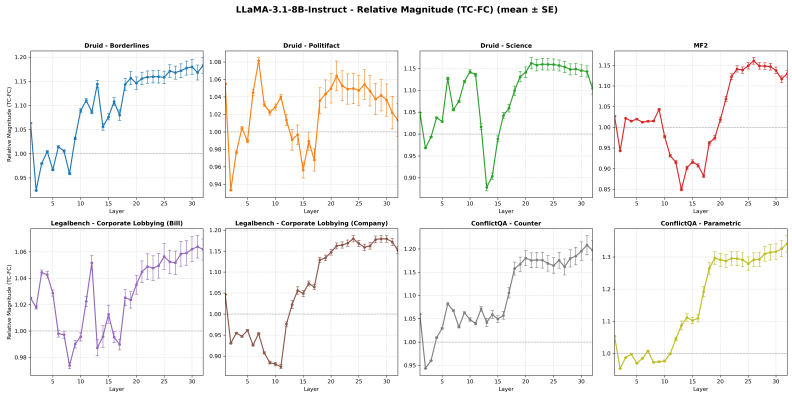


(c)

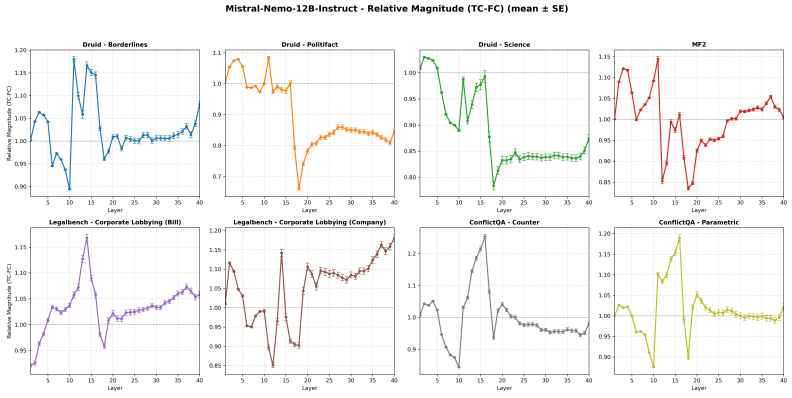


(d)

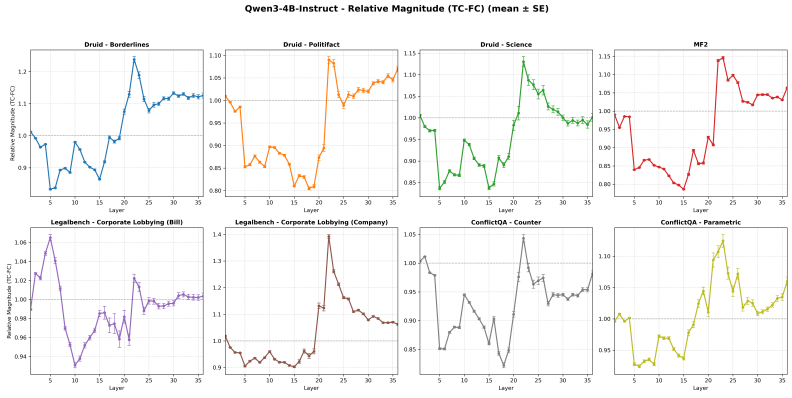
Figure 11: Layer wise plot of average θ in degrees across different models and datasets indicating the directional change in truth vectors when context is added. The error bars denote the standard error of mean



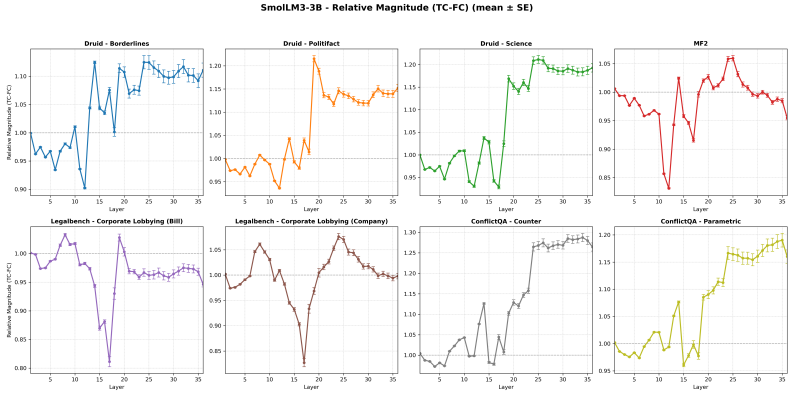
(a)



(b)



(c)



(d)

Figure 12: Layer wise plot of average relative magnitude across different models and datasets indicating the directional change in truth vectors when context is added. The error bars denote the standard error of mean