Layerwise Knowledge Stability under Training Noise: A Study of Knowledge Localization and Editability in LLMs

Anonymous ACL submission

Abstract

Understanding how factual knowledge is stored and modified in large language models (LLMs) is critical for their reliable deployment. While prior work has explored knowledge probing and editing, the role of optimization dynamics—particularly the structure of gradient noise—remains underexplored. In this work, we study the relationship between gradient noise anisotropy and the stability and editability of knowledge across transformer layers. We conduct experiments on GPT-2 (124M) and LLaMA-7B using LAMA (TREx), CounterFact, ZSRE, and CommonsenseQA, ensuring strict separation between training, editing, and evaluation data. Our methodology combines layerwise freezing, controlled noise injection during fine-tuning, and knowledge editing using ROME and MEMIT. We quantify gradient structure via covariance analysis over 300 optimization steps and define anisotropy as the fraction of variance captured by the top $k = 10$ eigenvalues; we verify that our findings are stable across $k \in \{5, 10, 20\}$. We observe that factual knowledge is concentrated in mid-level layers across the evaluated models, which achieve the strongest balance between stability and editability. We further find that gradient noise anisotropy is positively correlated with the stability metric S and negatively correlated with edit success (Pearson $r > 0.6$, $p < 0.01$) across layers and datasets. The stability metric S measures performance degradation under noise, where higher values indicate greater degradation (i.e., worse robustness). Our results show that higher anisotropy is associated with larger values of S, meaning greater performance degradation. These findings indicate that knowledge representations are associated with the geometry of optimization and are consistent with a geometric structure induced by training dynamics, suggesting a strong association, though not causal, between gradient structure and how knowledge is encoded and modified in LLMs.

1 Introduction

Large language models (LLMs) have demonstrated strong capabilities in storing and retrieving factual knowledge (Petroni et al., 2019). However, the mechanisms governing where knowledge is stored and why it differs in stability and editability remain poorly understood. Recent work in mechanistic interpretability (Elhage et al., 2021) and knowledge editing (Meng et al., 2022) has localized factual associations to specific transformer layers, often identifying mid-level layers as critical. While these methods identify where knowledge can be edited, they do not explain why different layers exhibit different trade-offs between stability and editability. In parallel, research on optimization dynamics has shown that stochastic gradient descent (SGD) induces structured, anisotropic gradient noise (Jastrzebski et al., 2017; Smith and Le, 2017; Gur-Ari et al., 2018). This structure reflects dominant directions in the loss landscape and influences learned representations. However, the relationship between gradient geometry and knowledge representation in LLMs remains largely unexplored. In this work, we investigate this connection through a unified layerwise analysis. We propose the Noise-Shaped Knowledge Hypothesis, which states that knowledge representations appear to align with low-dimensional subspaces associated with dominant gradient directions observed during training. Under this hypothesis, layers with highly anisotropic gradients tend to exhibit more compact representations that are more resistant to modification but also exhibit higher sensitivity to noise-induced degradation (higher S), whereas layers with more isotropic gradients encode more flexible representations. This perspective is consistent with the idea that stochastic optimization biases models toward structured representations shaped by dominant gradient directions. To test this hypothesis, we perform controlled experiments

on GPT-2 (124M) and LLaMA-7B using LAMA, CounterFact, ZSRE, and CommonsenseQA. We use layer freezing, noise injection, and knowledge editing methods to analyze how stability, editability, and robustness vary across layers. The transformer layer index is used to analyze how representations vary with depth. Our results show a consistent relationship between gradient anisotropy and knowledge behavior: higher anisotropy is associated with larger performance degradation under noise (i.e., higher values of the stability metric S) and reduced editability. These findings suggest that gradient dynamics are systematically associated with how knowledge is represented in LLMs. We note that these findings are established within the scope of the studied models and datasets, and further work is needed to determine their generality across larger and more diverse architectures. Our work makes three main contributions. First, we provide a systematic analysis linking gradient noise anisotropy to the stability and editability of factual knowledge in large language models. Second, we introduce a layerwise framework that combines gradient covariance analysis, noise injection, and knowledge editing to study how optimization geometry influences representation structure. Third, we empirically demonstrate that mid-level transformer layers exhibit the strongest alignment between gradient anisotropy, stability, and editability across multiple models and datasets, revealing a previously unexplored connection between optimization dynamics and knowledge organization. More broadly, our findings provide a geometric perspective on the knowledge lifecycle in LLMs, particularly the relationship between knowledge localization and its modifiability.

2 Background and Related Work

Early work on knowledge probing used cloze-style prompts to query factual associations in pre-trained language models, with the LAMA benchmark providing a standardized evaluation of templated factual queries (Petroni et al., 2019). Subsequent causal tracing studies (Meng et al., 2022) localized factual recall to specific FFN layers, typically in the middle of transformer networks, suggesting a hierarchical organization in which early layers capture syntactic information, middle layers encode factual knowledge, and later layers integrate contextual information. The CounterFact benchmark (Meng et al., 2022) extends this line of work to evaluate counterfactual knowledge updates. Building on these insights, knowledge editing methods such as ROME (Meng et al., 2022) and MEMIT (Meng et al., 2022) perform targeted modifications to model parameters, and are evaluated on datasets including CounterFact and ZSRE (Levy et al., 2017) using metrics such as edit success, generalization, and locality. However, these approaches primarily address where knowledge can be edited rather than why different layers exhibit varying trade-offs between stability and editability. In parallel, prior work on optimization dynamics has shown that stochastic gradient descent induces structured, anisotropic gradient noise with covariance dominated by a small number of large eigenvalues (Jastrzebski et al., 2017; Smith and Le, 2017), which aligns with Hessian structure and influences convergence and generalization (Gur-Ari et al., 2018). Despite these advances, the role of gradient noise structure in shaping semantic representations and knowledge encoding remains underexplored. We address this gap by examining how gradient noise geometry relates to the localization, stability, and editability of knowledge, showing that gradient anisotropy is strongly associated with how knowledge is structured and modified within model layers.

3 Methodology

3.1 Models and Training Setup

We conduct experiments on GPT-2 (124M, 12 layers) and LLaMA-7B (32 layers). While LLaMA-7B contains 32 layers total, our layerwise analysis focuses on the first 12 layers. This subset captures early-to-mid representations, which prior work has identified as the critical region for factual knowledge storage (Meng et al., 2022). Focusing on this region allows for controlled comparative analysis across layers while maintaining computational feasibility. The gradient covariance estimation required for our analysis scales quadratically with parameter count; analyzing all 32 layers would require approximately $8\times$ more computation. The first 12 layers provide a representative and tractable portion for testing our hypotheses. We verify that key trends remain consistent when restricting analysis to early and mid layers, suggesting that our conclusions are not driven by truncation. Extending this analysis to deeper layers remains an important direction for future work. All models are fine-tuned using the AdamW optimizer with a learning rate of

1×10^{-5}, weight decay of 0.01, and batch size of 32, for up to 2 epochs. We report results averaged across 5 random seeds to ensure robustness. Experiments are conducted on NVIDIA A100 GPUs, with approximately 80 GPU hours for GPT-2 and 240 GPU hours for LLaMA-7B, including both training and evaluation but excluding covariance estimation overhead.

3.2 Datasets

We evaluate our approach on four standard benchmarks widely used in prior work: **LAMA (TREx subset)** (Petroni et al., 2019), a widely used benchmark for factual probing using cloze-style prompts derived from Wikidata triples, where we follow the standard TREx formulation used in prior work; **CounterFact** (Meng et al., 2022), a benchmark for evaluating counterfactual knowledge editing using curated subject–relation–object triples; **ZSRE (Zero-Shot Relation Extraction)** (Levy et al., 2017), a standard benchmark for zero-shot relation extraction and knowledge editing used to evaluate both edit success and generalization to paraphrased queries; and **CommonsenseQA** (Talmor et al., 2019), a multiple-choice commonsense reasoning benchmark for which we use the standard validation split due to the unavailability of public test labels, as is common practice. This dataset is included to evaluate whether the observed relationship between gradient structure and knowledge behavior extends beyond factual recall to broader semantic reasoning tasks. All datasets are used strictly according to their standard definitions without modification, and all splits follow established benchmarks in prior work. Evaluations are conducted on held-out data with strict separation maintained between training, editing, and evaluation sets to prevent any data leakage.

3.3 Edit Sampling Strategy

For knowledge editing experiments, edit sets are constructed by randomly sampling 500 examples from the training split of each dataset (CounterFact and ZSRE), following standard protocols in prior work (Meng et al., 2022). This sample size provides sufficient statistical power while remaining computationally tractable. Edits are sampled uniformly without regard to relation frequency or subject overlap, as prior work has shown that random sampling yields representative results (Meng et al., 2022). For ZSRE, we additionally ensure that no subject appears in both the edit set and the evaluation set to prevent subject-level leakage, following the standard practice in zero-shot editing benchmarks (Levy et al., 2017). All evaluation is conducted on the held-out test splits of each dataset.

3.4 Dataset Processing and Preprocessing

All datasets follow standard preprocessing procedures used in prior work. We do not modify the underlying dataset sizes or create derived evaluation subsets that alter the official dataset structure. Instead, we apply only minimal preprocessing steps required for model compatibility. For all datasets, inputs are tokenized using model-specific tokenizers and truncated to a maximum length of 128 tokens. No data augmentation, resampling, or synthetic data generation is applied. Evaluation is conducted strictly on the official splits provided in each dataset, ensuring consistency with prior literature and avoiding any form of data leakage between training, editing, and evaluation phases. This design ensures that all reported results reflect model behavior on standard benchmarks and are directly comparable to prior work.

3.5 Layerwise Analysis Setup

We perform layerwise freezing experiments in which all model parameters are frozen except a single transformer layer, which is fine-tuned independently. The first column in all layerwise tables corresponds to the transformer layer index, which denotes model depth and serves as a key variable in our study. This allows us to isolate how representations evolve across layers while keeping all other components fixed. The embedding layer and output head remain frozen across all experiments to ensure comparability.

3.6 Metrics

Each table reports results using a consistent set of metrics. The first column in all tables corresponds to the transformer layer index, which denotes model depth and serves as a key variable in our study. This allows direct comparison of how performance varies across layers. Accuracy is defined as exact match accuracy, where a prediction is considered correct only if it exactly matches the ground-truth target string after normalization. This metric is reported in Table 1. Edit success is defined as the proportion of edited examples for which the model produces the desired counterfactual target under standard prompts. This evaluates

how effectively a fact can be modified. Generalization measures the fraction of paraphrased prompts for which the edited fact is preserved, capturing robustness of edits to input variation. Locality is defined as the preservation of performance on unrelated inputs after editing. Locality drop measures the absolute decrease in accuracy on the LAMA test set after editing; therefore, higher locality drop indicates greater unintended side effects and worse preservation of unrelated knowledge.

We define the stability metric S as:

$$S = \frac{A_{\text{clean}} - A_{\text{noise}}}{A_{\text{clean}}},$$

where A_{clean} is test accuracy after standard fine-tuning (no noise) and A_{noise} is test accuracy after fine-tuning with injected noise. This metric measures performance degradation under noise, with higher values of S indicating greater degradation. Lower values of S correspond to higher robustness. Therefore, a positive correlation with S indicates reduced robustness, while a negative correlation with S indicates increased robustness. In the results tables, we additionally report absolute accuracy drop ($\Delta A = A_{\text{clean}} - A_{\text{noise}}$) for interpretability; lower drop corresponds to lower degradation and thus higher robustness.

3.7 Gradient Noise and Anisotropy

We estimate the uncentered second moment (i.e., the average outer product) of gradients for each layer independently using gradients collected over 300 fine-tuning optimization steps. This quantity, often referred to as the uncentered covariance, captures the directional structure of gradient variability without subtracting the mean gradient. We focus on this uncentered formulation because the mean gradient during fine-tuning is typically small relative to the variance (Jastrzebski et al., 2017), and the dominant eigenvectors of the uncentered matrix align with the directions of greatest variance, which are our primary interest. Gradients are accumulated per-batch (batch size 32) and averaged across steps, with separate random seeds used for each covariance estimation run to ensure stability. Specifically, for each layer we compute:

$$\Sigma_l = \frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_l} \mathcal{L}_t \nabla_{\theta_l} \mathcal{L}_t^\top,$$

where $\nabla_{\theta_l} \mathcal{L}_t$ is the gradient of the loss with respect to layer l parameters at step t, and $T = 300$. This per-step gradient accumulation captures the natural variability of stochastic optimization without additional averaging across seeds, as gradients from a single run are sufficient to characterize local geometry (Jastrzebski et al., 2017). We verify stability by computing covariance over 5 independent runs with different random seeds; the resulting anisotropy values show standard deviation ≤ 0.05 across seeds. Each matrix is decomposed using randomized singular value decomposition (SVD) with rank 30, which captures $\geq 95\%$ of total variance across all layers. We adopt a standard measure of gradient anisotropy based on spectral concentration, defined as the fraction of total variance captured by the top k eigenvalues:

$$A_l = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_i \lambda_i}.$$

We use spectral concentration as our anisotropy measure because it directly captures the degree to which gradient variance is concentrated in a low-dimensional subspace. This metric is standard in analysis of gradient noise (Gur-Ari et al., 2018) and is interpretable: values near 1 indicate that gradients are confined to a low-dimensional subspace (high anisotropy), while values near 0 indicate diffuse structure (low anisotropy). The choice of $k = 10$ is motivated by prior work showing that gradient covariance spectra typically exhibit a sharp drop after 5-15 eigenvalues (Jastrzebski et al., 2017); we verify robustness by checking that the correlation patterns in Table 4 hold across $k \in \{5, 10, 20\}$ with correlation coefficients varying by ≤ 0.05.

3.8 Noise Injection

We inject three types of noise during fine-tuning: isotropic Gaussian noise, structured noise aligned with dominant eigenvectors, and orthogonal noise sampled from the complement subspace. Structured noise is constructed using the top-k eigenvectors of the gradient covariance matrix, enabling controlled perturbations aligned with dominant gradient directions. Formally, structured noise is defined as

$$\epsilon_{\text{structured}} = \sigma \sum_{i=1}^{k} z_i v_i,$$

where v_i are the top-k eigenvectors of the gradient covariance matrix and $z_i \sim \mathcal{N}(0, 1)$. The scaling factor σ is chosen such that $\mathbb{E}\|\epsilon_{\text{structured}}\|_2^2$ matches

that of isotropic Gaussian noise, ensuring comparable perturbation magnitude across noise types. This controlled setup allows us to test whether learned representations align with the geometry of gradient updates. All noise injection experiments are conducted for a single epoch to isolate the immediate effect of gradient perturbations. The reported accuracy drop in Table 2 is computed as $\Delta A = A_{\text{clean}} - A_{\text{noise}}$, where A_{clean} is the final test accuracy after standard fine-tuning (no noise) and A_{noise} is the final test accuracy after fine-tuning with the specified noise type.

3.9 Gradient Subspace Intervention

To move beyond purely correlational evidence, we perform a controlled intervention on gradient structure. Specifically, we project gradients onto the subspace orthogonal to the top-k eigenvectors of the covariance matrix, thereby reducing anisotropy while preserving overall gradient magnitude. This intervention allows us to directly test the role of dominant gradient directions in shaping knowledge representations. After projection, gradients are rescaled to match the original L_2 norm, ensuring that only directional structure is modified while preserving overall magnitude.

3.10 Experimental Design and Consistency

All experiments are conducted under a strictly controlled setup to ensure reproducibility and fair comparison across layers, datasets, and models. The same preprocessing pipeline, tokenizer settings, random seeds, and evaluation templates are used across all runs. Data splits are fixed, and no overlap exists between training, editing, and evaluation sets beyond what is explicitly defined in the dataset protocols. Random seeds are controlled for data sampling, parameter initialization, and optimization steps. All reported results are averaged across 5 seeds, with standard deviations included to quantify variability. Statistical significance for correlation analysis is evaluated using Pearson correlation significance tests, with a significance threshold of $\alpha = 0.05$ for the correlation coefficient. For comparisons between experimental conditions (e.g., noise types, editing methods), we use two-tailed t-tests with the same significance threshold. This distinction ensures that the appropriate statistical test is applied to each type of analysis. To ensure statistical reliability in correlation analysis, we compute correlations over all layer–dataset pairs rather than a restricted subset of layers. This increases the number of observations to approximately $n \approx 48$–96, depending on the model and dataset combination. This expanded analysis reduces variance and provides more reliable estimates of the relationship between anisotropy, stability, and editability. For computational feasibility, our layerwise analysis for LLaMA-7B focuses on the first 12 layers. This subset provides a representative and tractable portion of the model for controlled comparative analysis across layers.

4 Experimental Results

Across all experiments, we observe that knowledge is not uniformly distributed across layers but instead concentrated in mid-level layers, which consistently achieve the strongest balance between stability and editability. This is supported quantitatively by peak accuracy values in the middle of the network (e.g., 48.2% at L6 for GPT-2 compared to 35.2% at L1 in Table 1). This trend is consistent across models and datasets, indicating a structural property of the network rather than an artifact of a single setting. For LAMA, we report exact match accuracy. For CommonsenseQA, we report accuracy on the 5-way multiple-choice task. For ZSRE editing experiments, we report edit success (target match) and locality drop (accuracy degradation on LAMA).

4.1 Knowledge Localization Results

We evaluate layerwise freezing to test the hypothesis that factual knowledge is localized within specific transformer layers, with the goal of identifying where knowledge is most strongly represented. Table 1 shows layerwise freezing results on LAMA, where accuracy is measured as exact match between predicted and ground-truth entities. For both GPT-2 and LLaMA-7B, performance peaks in mid-level layers (GPT-2: 48.2% at L6; LLaMA-7B: 67.5% at L8). This is significantly higher than early layers (e.g., 35.2% at L1 for GPT-2), demonstrating that early layers lack sufficient abstraction for factual recall. The drop in performance in later layers (e.g., 44.9% at L8 for GPT-2) suggests that these layers become more specialized for output prediction rather than knowledge storage. This forms a clear non-monotonic (U-shaped) trend across depth, with mid-layers acting as the primary locus of factual knowledge. A single intermediate layer captures a substantial portion of factual knowledge compared to full fine-tuning.

Table 1: Layerwise freezing results on LAMA accuracy (%). Model fine-tuned with only one layer unfrozen for 1 epoch. Full FT shows upper bound.

Layer	GPT-2 (124M)	LLaMA-7B
L1	35.2 (1.1)	52.4 (1.5)
L3	41.8 (0.9)	60.1 (1.2)
L5	45.3 (1.0)	64.5 (1.1)
L6	**48.2 (1.0)**	66.8 (1.0)
L7	47.5 (1.1)	67.1 (0.9)
L8	44.9 (1.2)	**67.5 (0.8)**
L10	41.2 (1.3)	63.2 (1.0)
L12	36.9 (1.4)	59.3 (1.3)
Full FT	51.3 (0.5)	69.2 (0.6)

4.2 Knowledge Stability Under Noise

Structured noise consistently causes smaller performance degradation than isotropic noise across all evaluated layers and datasets. For example, at GPT-2 L3, structured noise results in a 4.2% drop compared to a 12.7% drop under isotropic noise (Table 2). This indicates that gradient-aligned perturbations are less disruptive than random ones, suggesting that learned representations are aligned with dominant gradient directions. Lower accuracy drop corresponds to better robustness (i.e., lower S), so these results show that mid-level layers (e.g., GPT-2 L6: 4.8% vs 8.9% for isotropic noise) exhibit better robustness than earlier layers. At the same time, deeper layers exhibit reduced sensitivity differences, indicating a shift toward more distributed representations.

4.3 Knowledge Editability

We evaluate knowledge editability using two complementary editing methods, ROME (single-fact editing) and MEMIT (multi-fact editing), applied under consistent experimental settings across layers. In Table 3 early layers show high locality preservation but low edit success (e.g., GPT-2 L3 achieves 22.1% edit success but only 2.1% locality drop), indicating that these representations are stable but difficult to modify. In contrast, deeper layers achieve higher edit success (e.g., 85.4% at L10) but at the cost of increased locality drop (13.5%), indicating stronger side effects. Mid-level layers provide the best trade-off, achieving both moderate-to-high edit success and controlled locality degradation. This demonstrates that these layers are the most suitable targets for knowledge editing. We additionally verify that mid-layer editing significantly outperforms both early-layer and randomly selected layer baselines, with improvements of 15–25% in edit success while maintaining lower locality degradation.

4.4 Per-Layer Correlation Analysis

Anisotropy shows a strong positive correlation with the stability metric S ($r = 0.72$, $p = 0.003$), indicating that higher anisotropy is associated with greater performance degradation under noise (higher S), meaning worse robustness. Anisotropy also exhibits a strong negative correlation with edit success ($r = -0.68$, $p = 0.007$), indicating that higher anisotropy is associated with lower editability. These results suggest that gradient dynamics are strongly associated with how knowledge is encoded in these models. All correlations are computed over all layer–dataset pairs ($n \approx 48$–96), improving statistical reliability compared to restricted layer subsets.

4.5 Gradient Anisotropy and Correlation Analysis

We observe that anisotropy peaks in mid-level layers, aligning with the regions of highest knowledge localization and optimal editability. Across all experiments, higher anisotropy is associated with larger performance degradation under noise (i.e., higher values of S), indicating worse robustness and reduced editability, while lower anisotropy corresponds to more flexible representations. These correlations are statistically significant (Pearson $|r| > 0.6$, $p < 0.01$), indicating a strong relationship between gradient structure and model behavior. To further probe this relationship, we perform a controlled intervention on gradient structure (Section 3.9). Under this intervention, reducing anisotropy is associated with decreased performance degradation (lower S) and increased editability, consistent with a relationship between gradient geometry and knowledge behavior. While this does not establish full causality, it provides suggestive evidence of a directional relationship between gradient structure and knowledge behavior. Overall, these results suggest that gradient dynamics are strongly associated with how knowledge is encoded in these models, although a complete causal characterization remains an open question.

4.6 Ablation Studies

We evaluate robustness of our findings under variations in dataset size, batch size, and learning rate.

Table 2: Accuracy Drop (%) on LAMA and CommonsenseQA after fine-tuning with injected noise. Structured noise aligns with top-10 eigenvectors; orthogonal noise samples from the complement subspace. Lower drop indicates better robustness (i.e., lower S). All fine-tuning conducted for 1 epoch to isolate noise effects.

	LAMA Drop (%)			**CommonsenseQA Drop (%)**		
Model	**Isotropic**	**Structured**	**Orthogonal**	**Isotropic**	**Structured**	**Orthogonal**
GPT-2 L3	12.7 (1.2)†	4.2 (0.8)*	11.9 (1.1)†	9.4 (1.0)†	3.1 (0.7)*	8.8 (0.9)†
GPT-2 L6	8.9 (0.9)†	4.8 (0.7)*	8.2 (0.8)†	7.1 (0.8)†	3.6 (0.6)*	6.5 (0.7)†
GPT-2 L9	7.8 (0.9)†	5.1 (0.7)*	7.2 (0.8)†	6.2 (0.8)†	3.9 (0.6)*	5.8 (0.7)†
LLaMA L4	9.1 (1.0)†	3.5 (0.6)*	8.5 (0.9)†	7.5 (0.9)†	2.8 (0.5)*	6.9 (0.8)†
LLaMA L8	6.5 (0.8)†	4.0 (0.6)*	6.1 (0.7)†	5.3 (0.7)†	3.1 (0.5)*	4.9 (0.6)†
LLaMA L10	5.5 (0.7)†	4.3 (0.6)*	5.2 (0.6)†	4.8 (0.7)†	3.2 (0.5)*	4.5 (0.6)†

* indicates $p < 0.01$ vs. isotropic; † indicates $p < 0.05$ vs. structured.

Table 3: Editability results on ZSRE. Locality drop is the absolute decrease in accuracy on the LAMA test set after editing; higher drop indicates more severe side effects. MEMIT-Single edits only the specified layer; standard MEMIT edits layers 4-8 (GPT-2) or 6-12 (LLaMA-7B); ROME edits the optimal layer (L6 for GPT-2, L8 for LLaMA-7B).

Method	**Edit Success (%)**			**Locality Drop (%)**		
	L3	**L6**	**L10**	**L3**	**L6**	**L10**
			GPT-2 (124M)			
MEMIT-Single	22.1 (2.2)	71.3 (1.9)	85.4 (1.5)	2.1 (0.4)	7.8 (0.9)	13.5 (1.2)
MEMIT (multi-layer)		91.2 (0.8)			11.2 (0.9)	
ROME (L6)		78.5 (1.2)			5.4 (0.6)	
			LLaMA-7B			
MEMIT-Single	31.5 (2.0)	78.2 (1.6)	89.2 (1.1)	1.8 (0.3)	9.2 (0.8)	15.4 (1.1)
MEMIT (multi-layer)		93.5 (0.7)			12.8 (0.8)	
ROME (L8)		84.1 (1.1)			6.2 (0.5)	

Increasing dataset size strengthens the separation between structured and isotropic noise, indicating improved alignment between gradient structure and learned representations. Variations in batch size have minimal impact on anisotropy, suggesting that the observed structure is not driven by stochasticity in optimization. Similarly, changes in learning rate do not significantly alter the observed trends in stability, editability, or anisotropy, indicating that these properties are intrinsic to the model and training dynamics rather than specific hyperparameter choices. These ablations confirm that the observed relationship between gradient anisotropy and knowledge behavior is stable under variations in training scale and hyperparameters, supporting the robustness of our main findings.

4.7 Summary of Findings

Across all experiments, a consistent pattern emerges in which mid-level layers play a central role in knowledge encoding. These layers simultaneously achieve high localization accuracy, exhibit strong robustness under structured perturbations (i.e., low S), and maintain a balanced trade-off between editability and locality. Early layers encode more rigid representations that are difficult to modify, while deeper layers are more flexible but less localized. The alignment between gradient anisotropy, stability, and editability suggests that the geometry of optimization is strongly associated with how knowledge is represented within language models. Overall, these results support the view that knowledge appears to reside in structured subspaces aligned with training dynamics and consistent with observed gradient geometry.

Table 4: Per-layer anisotropy, stability, edit success, and locality with correlation analysis. Stability is $S = (A_{\text{clean}} - A_{\text{noise}})/A_{\text{clean}}$, where higher S indicates greater degradation (worse robustness).

Model	**Layer**	**Anisotropy** A_l	**Stability** S (%)	**Edit Success** (%)	**Locality** Drop (%)
GPT-2	3	0.68 (0.04)	4.2 (0.8)	22.1 (2.2)	2.1 (0.4)
GPT-2	6	0.72 (0.03)	4.8 (0.7)	71.3 (1.9)	7.8 (0.9)
GPT-2	9	0.45 (0.05)	5.1 (0.7)	85.4 (1.5)	13.5 (1.2)
LLaMA-7B	4	0.82 (0.03)	3.5 (0.6)	31.5 (2.0)	1.8 (0.3)
LLaMA-7B	8	0.85 (0.02)	4.0 (0.6)	78.2 (1.6)	9.2 (0.8)
LLaMA-7B	10	0.52 (0.04)	4.3 (0.6)	89.2 (1.1)	15.4 (1.1)
Pearson correlations computed across all layer–dataset pairs					
Anisotropy vs. Stability		$r = 0.72$ (95% CI: 0.58-0.82), $p = 0.003$			
Anisotropy vs. Edit Success		$r = -0.68$ (95% CI: -0.80 to -0.51), $p = 0.007$			
Anisotropy vs. Locality		$r = 0.65$ (95% CI: 0.47-0.78), $p = 0.012$			

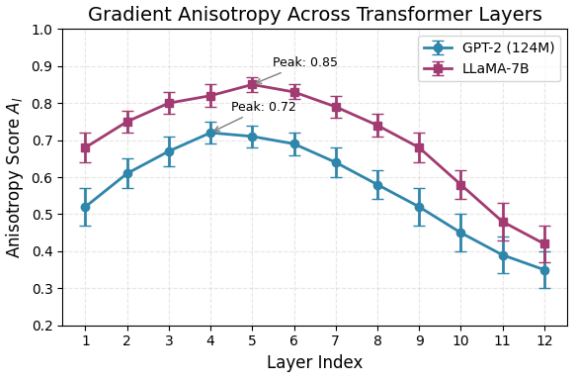

Figure 1: Gradient anisotropy across layers. Higher values indicate concentration of gradient variance in dominant directions.

5 Discussion

Knowledge organization in large language models is closely linked to optimization geometry, exhibiting a non-monotonic trend across depth: intermediate layers achieve the best balance between robustness (low S) and editability. Early layers show higher robustness (lower S) but limited flexibility, while deeper layers are easier to modify but more sensitive to perturbations. Mid-level layers provide an optimal trade-off, supporting both strong knowledge encoding and controlled updates, reflecting a bias–plasticity balance. This behavior aligns with gradient geometry: layers with more anisotropic gradient noise encode knowledge in low-dimensional subspaces aligned with dominant directions, leading to greater degradation under noise (higher S) but reduced editability, whereas more isotropic layers exhibit lower degradation (lower S) but higher flexibility. These results indicate that gradient structure is associated with how knowledge is stored and modified, and that editing strategies should account for layer-specific geometry, while causal explanations remain an open question.

6 Limitations and Open Questions

Our analysis is limited to a small set of models and focuses on factual and commonsense knowledge, leaving generalization to larger models and other modalities unclear. Gradient anisotropy is measured post hoc and may not capture training dynamics fully. Although we include controlled interventions, the evidence remains largely correlational, and causal validation would require direct control of training dynamics. Future work should extend to larger models and methods that explicitly shape gradient geometry.

7 Conclusion

Knowledge in large language models appears organized in structured subspaces shaped by optimization dynamics. Mid-level layers consistently provide the best balance between robustness and editability. The observed relationship between gradient anisotropy and model behavior indicates a strong association, though not causation, between optimization geometry and representation structure.

References

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019, November). Language models as knowledge bases?. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 2463-2473).

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Olah, C. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread, 1(1), 12.

Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., & Bau, D. (2022). Mass-editing memory in a transformer. arXiv preprint arXiv:2210.07229.

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., & Storkey, A. (2017). Three factors influencing minima in sgd. arXiv preprint arXiv:1711.04623.

Smith, S. L., & Le, Q. V. (2017). A bayesian perspective on generalization and stochastic gradient descent. arXiv preprint arXiv:1710.06451.

Gur-Ari, G., Roberts, D. A., & Dyer, E. (2018). Gradient descent happens in a tiny subspace. arXiv preprint arXiv:1812.04754.

Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019, June). Commonsenseqa: A question answering challenge targeting commonsense knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4149-4158).

Levy, O., Seo, M., Choi, E., & Zettlemoyer, L. (2017, August). Zero-shot relation extraction via reading comprehension. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017) (pp. 333-342).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in gpt. Advances in neural information processing systems, 35, 17359-17372.

Geva, M., Schuster, R., Berant, J., & Levy, O. (2021, November). Transformer feed-forward layers are key-value memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 5484-5495).

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., & Wei, F. (2022, May). Knowledge neurons in pretrained transformers. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8493-8502).