CLARE-ty Amid Chaos: Quantifying Representational Entanglement to Predict Ripple Effects in LLM Editing

Anonymous ACL submission

Abstract

The static knowledge representations of large language models (LLMs) inevitably become outdated or incorrect over time. While model-editing techniques offer a promising solution by modifying a model's factual associations, they often produce unpredictable ripple effects, which are unintended behavioral changes that propagate even to the hidden space. In this work, we introduce CLARE, a lightweight representation-level technique to identify where these ripple effects may occur. Unlike prior gradient-based methods, CLARE quantifies entanglement between facts using forward activations from a single intermediate layer, avoiding costly backward passes. To enable systematic study, we prepare and analyse a corpus of 11,427 facts drawn from three existing datasets. Using CLARE, we compute large-scale entanglement graphs of this corpus for multiple models, capturing how local edits propagate through representational space. These graphs enable stronger preservation sets for model editing, audit trails, efficient red-teaming, and scalable post-edit evaluation. In comparison to baselines, CLARE achieves an average of 62.2% improvement in Spearman correlation with ripple effects while being 2.74× faster, and using 2.85× less peak GPU memory. Besides, CLARE requires only a fraction of the storage needed by the baselines to compute and preserve fact representations. Our entanglement graphs and corpus are available at https://anonymous.4open.science/r/CLaRE-488E.

1 Introduction

Large language models (LLMs) are increasingly used in applications such as search (Zhao et al., 2023), dialogue systems (Raza et al., 2025), education (Yigci et al., 2025), and decision support (Yang et al., 2023). However, their encoded knowledge needs updating, as they may contain misinformation or outdated content (Sallami et al., 2024; Lin et al., 2025). The fine-tuning process updates model parameters but scales poorly for large models (Fang et al., 2025). Retrieval-augmented generation (RAG) supplements the LLM's input with external context, but may fail when it misinterprets the retrieved content or when the retrieved content itself is inaccurate (Agrawal et al., 2024). As a complementary strategy, model editing updates specific factual associations in model weights, but may have its own trade-offs.

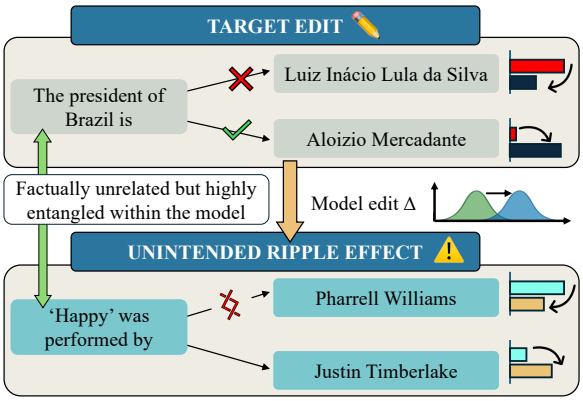


Figure 1: A targeted update to a political fact may inadvertently alter the model's prediction for an unrelated musical fact, despite no semantic connection. This demonstrates how edits can trigger ripple effects far beyond the intended factual neighborhood.

Parameter-modifying model-editing techniques (Meng et al., 2023, 2022) localize and alter weights for targeted factual associations. Although they offer significant improvements in computational efficiency compared to fine-tuning, they also lead to ripple effects. These are unintended behavioral changes at other, semantically related (or even unrelated) model outputs (Cohen et al., 2024). While some ripple effects may enhance generalization, many introduce new hallucinations, distort factual consistency, or

degrade model performance. Prior efforts such as GradSim (Qin et al., 2024b) and the RippleEdits benchmark (Cohen et al., 2024) have attempted to quantify how these effects propagate to semantically neighboring facts, typically within one or two hops of the edited entity in a knowledge graph. They reveal that LLM representations are highly interconnected: altering one fact may influence others that share representational subspaces. However, ripple effects can also extend to hidden spaces, producing distortions even where no direct factual relationship exists (Wang et al., 2025), as illustrated in Figure 1. These *inter-domain* ripple effects may degrade a model's reliability and deteriorate its performance. Furthermore, although GradSim estimates entanglement between two facts, we show in Section 5 that such gradient similarity poorly correlates with inter-domain ripple effects, and it is computationally inefficient as it calculates the entire gradient for each fact.

These cross-domain distortions due to model editing remain poorly understood and largely unmeasured in current evaluation frameworks. For comprehensive analysis, we curate 11,427 facts spanning diverse domains, drawn from 3 existing datasets (Zhong et al., 2023; Cohen et al., 2024; Liu et al., 2025). This corpus, containing 212 unique prompt formats and 6,140 unique subjects, enables analysis of how model edits propagate globally through a model's knowledge base and how these dynamics vary across model architectures (Section 5). We introduce **CLARE** (Critical Layer Representation Entanglement), a lightweight representation-level technique for identifying where ripple effects are most likely to occur (Section 4). Unlike computationally heavy gradient-based approaches like GradSim, CLARE quantifies entanglement directly from forward activations from a single intermediate layer, without requiring backward passes or gradient computation. Our experiments show that in comparison to GradSim, CLARE achieves an average of 62.2% improvement in Spearman correlation with observed ripple effects across multiple editing techniques and models (Section 5.1). Beyond predictive accuracy, CLARE is 2.74× faster and uses 2.85× lower peak GPU memory, while requiring only a fraction of storage needed to preserve fact representations (Section 5.2).

We release entanglement graphs computed with CLARE over the corpus of 11,427 facts across multiple models, providing a basis for safer, more interpretable, and cost-effective model editing with practical utility for both research and deployment. These graphs enable safer editing by enabling stronger preservation-set construction and facilitate audit trails and post-edit explainability. They further support downstream applications such as budget-constrained red-teaming of editing techniques by identifying high-risk regions within the model (Section 5.3). Together, these capabilities help in assessing ripple effects and advance reliable, auditable and interpretable model editing.

Contributions. ❶ We introduce CLARE, a lightweight and scalable technique for identifying where ripple effects are most likely to occur. CLARE offers 2.74× speedup, and 2.85× lower peak GPU memory usage as compared to gradient-based methods. ❷ We curate 11,427 facts across diverse domains, to systematically study how edits propagate globally across a model's knowledge base. CLARE achieves up to 0.31 higher Spearman correlation with observed ripple effects, averaging a 62.2% improvement over gradient-based methods in predictive alignment. ❸ We release large-scale entanglement graphs for these 11,427 facts computed with CLARE for multiple models, enabling construction of stronger preservation sets for model editing, and support cost-effective red-teaming and post-edit evaluation at scale.

2 Related Work

Model-editing techniques. Parameter-modifying and parameter-preserving approaches are two primary paradigms for model editing in LLMs. Parameter-modifying techniques directly alter model weights through targeted optimization. Approaches like MEND use gradient decomposition for fast, local edits (Mitchell et al., 2022a). Recently, locate-then-edit techniques have emerged which target specific knowledge, yet they are also subject to ripple effects (Meng et al., 2023, 2022; Fang et al., 2025; Gu et al., 2024; Ma et al., 2025). Parameter-preserving techniques augment models with external modules instead of modifying weights (Mitchell et al., 2022b; Zheng et al., 2023; Huang et al., 2023; Hartvigsen et al., 2023; Chen et al., 2024). Emerging techniques integrate knowledge graphs and chain-of-thought reasoning with parameter modification (Zhang et al., 2024a; Dong et al., 2025; Zhao et al., 2024). We focus on parameter-modifying techniques, as they update the model parameters.

Model-editing evaluation techniques and benchmarks. CounterFact (Meng et al., 2022) and CounterFactPlus (Hoelscher-Obermaier et al., 2023) benchmark specificity through counterfactual edits. MQuAKE (Zhong et al., 2023) and ThinkEval (Baser et al., 2025) evaluate multi-hop and multi-step reasoning, while KnowEdit (Zhang et al., 2024b) provides a unified framework across six datasets. LEME (Rosati et al., 2024) and Evedit (Liu et al., 2024) target long-form and event-based editing. UnKEBench (Deng et al., 2025) tests editing over noisy, unstructured data, and ZsRE (Levy et al., 2017) focuses on QA-based factual samples. Know-MRI (Liu et al., 2025) offers a toolkit integrating 13 datasets to analyse LLM knowledge mechanisms. RippleEdits (Cohen et al., 2024) is a benchmark of factual edits meant to capture ripple effects caused by editing techniques. While RippleEdits focuses on semantically or graph-neighboring facts, our goal is to detect ripple effects in the hidden space, including across unrelated or cross-domain facts. Hence, we curate a corpus of 11,427 facts from MQuAKE, RippleEdits and Know-MRI, enabling broader domain coverage and fine-grained analysis.

Ripple Effects and Entanglement. DiKE (Zhang et al., 2025) is an editing technique which uses a disentanglement module to preserve other facts about the same entity, but does not consider damage brought to the hidden space. SIR (Wang et al., 2025) is an editing technique which addresses ripple effects in hidden space, but it only detects them after editing. It requires evaluation after editing followed by re-editing to repair the damage, making it reactive rather than preventive, and unsuitable for mapping entanglement pre-edit or at scale. GradSim (Qin et al., 2024b) uses gradient similarity to estimate entanglement, but computing full gradients for each fact remains inefficient for large corpora. These limitations motivate the need for alternatives like CLARE, which directly estimates entanglement using forward activations from a single intermediate layer. Through our experiments, we show that CLARE achieves higher predictive alignment with observed ripple effects while scaling efficiently to thousands of facts, enabling corpus-wide entanglement mapping.

3 Preliminary

As an input token $x_{[t]}$ moves through hidden layer l of a transformer, its representation is updated by residual, attention, and feedforward components:

$$h^l_{[t]}(x) = h^{l-1}_{[t]}(x) + a^l_{[t]}(x) + m^l_{[t]}(x), \qquad (1)$$

where $h^0_{[t]}(x)$ denotes the input embedding, $a^l_{[t]}$ is the attention output, and $m^l_{[t]}$ is the MLP (feedforward) contribution:

$$a^l_{[t]} = \mathrm{attn}\big(h^{l-1}_{[1]}, \ldots, h^{l-1}_{[t]}\big), \qquad (2)$$

$$m^l_{[t]} = W^l_{\mathrm{out}}\, \sigma\big(W^l_{\mathrm{in}}\, \gamma(h^{l-1}_{[t]})\big). \qquad (3)$$

Here, γ denotes layer normalization and σ a nonlinear activation. The residual stream $h^l_{[t]}$ aggregates attention and nonlinear feedforward signals, and this layer-wise buildup enables the LLM to compose and propagate factual associations to its final prediction. Prior works find that factual associations are often localized in a small subset of intermediate MLP layers, termed the critical MLP layers $\mathcal{R}$ (Meng et al., 2022, 2023). Intuitively, these layers perform an associative "key→value" computation: subject tokens elicit key-like activation patterns, and the MLPs generate values that later shape the output logits. Interventions at these layers may yield the largest changes in answers to factual prompts, identifying them as major causal sites for storing and retrieving facts.

A standard procedure to identify critical layers is causal tracing (Meng et al., 2022, 2023). Causal tracing holds the model computation fixed up to a candidate layer and then swaps or ablates subsequent activations from a reference input, measuring the downstream effect on the model's prediction. Layers whose interventions produce the largest change receive the highest causal scores and are labeled as the critical mediators for that fact. This per-layer causal ranking provides a principled basis to select layers for targeted editing. These mechanistic insights have motivated a family of parameter-modifying editing techniques, such as ROME (Meng et al., 2022), MEMIT (Meng et al., 2023) and EMMET (Gupta et al., 2024). (Gupta et al., 2025) characterize approximate linear associative structure in factual storage, (Yao et al., 2024) describe distributed "knowledge circuits" that implement higher-level computation , and (Guerin et al., 2025) analyse neuron-level distinctions between language-structural and semantic storage.

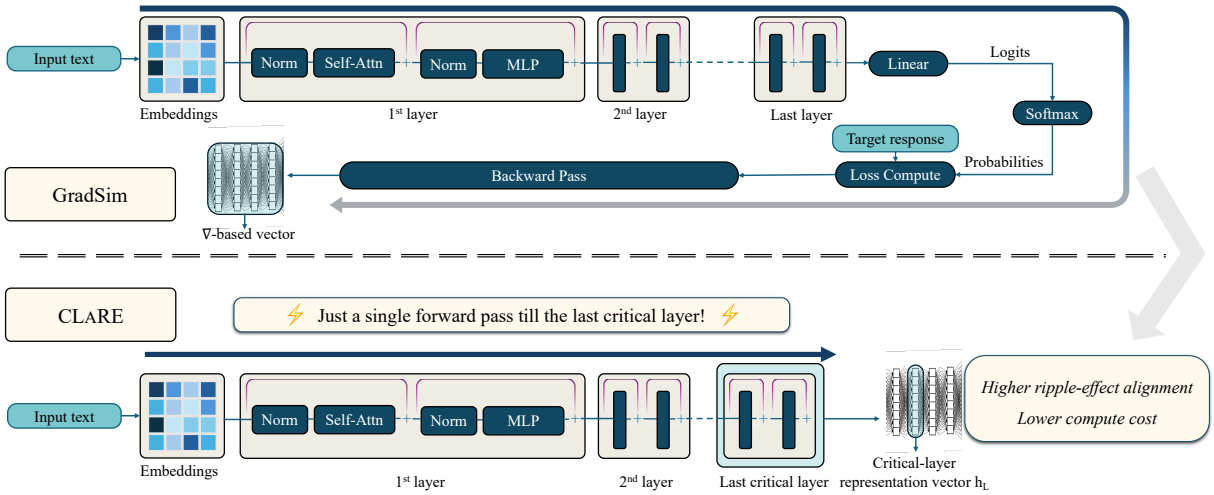


Figure 2: Comparison between GradSim and CLARE. For each fact, GradSim computes the entire gradient, while CLARE uses a single forward pass up till the last critical layer, enabling faster and scalable entanglement mapping.

4 CLARE: A lightweight and scalable technique for identifying ripple effects

We represent facts as triplets (subject s, relation r, object o), and the input prompt as $p_i = p(s_i, r_i)$. Let the memory to be updated in the model be (s_i, r_i, o_i). To simplify notation, following Eq. 1, let $h_i^l = h_{[S]}^l(p_i)$, where S is the last subject token of p_i. Let $\mathcal{L} = \max(\mathcal{R})$ be the last critical layer of the model. If a fact's causal influence is concentrated in a contiguous band of layers $\mathcal{R}$, then $\mathcal{L}$ tends to be the deepest point where the fact is still clearly represented before downstream mixing by later attention and MLPs. Using this layer as a probe gives a representation that contains the associative signal produced by earlier critical layers. As later layers can combine or diffuse information, a representation taken at $\mathcal{L}$ is more stable for cross-fact comparisons than either earlier activations or the final output logits that also encode decoding constraints. Thus, for a fact (s_i, r_i, o_i):

$$h_i^{\mathcal{L}} = h_i^0 + \sum_{l=1}^{\mathcal{L}} a_i^l + \sum_{l=1}^{\mathcal{L}} m_i^l \tag{4}$$

$h_i^{\mathcal{L}}$ is compact and informative, and relatively local snapshot of where the model "stores" the fact as compared to entire model gradients. To be specific, we do not perform weight edits. We use $h_i^{\mathcal{L}}$ purely as a probe. We quantify entanglement, i.e., whether two facts (s_i, r_i, o_i) and (s_j, r_j, o_j), overlap in the way model encodes them prior to any editing. This overlap serves as a proxy to identify which facts may get impacted by ripple effects.

$$\begin{aligned}\text{CLARE}(i,j) &= \cos\big(h_i^{\mathcal{L}}, h_j^{\mathcal{L}}\big) \\ &= \frac{\langle h_i^{\mathcal{L}}, h_j^{\mathcal{L}} \rangle + \varepsilon}{\|h_i^{\mathcal{L}}\| \, \|h_j^{\mathcal{L}}\| + \varepsilon}\end{aligned} \tag{5}$$

A high CLARE entanglement score indicates that the model stores both facts in similar subspaces, increasing the likelihood that an edit to one will influence the other. Each fact requires only a single forward pass up to $\mathcal{L}$ (and not the last layer) to obtain its hidden representation at $\mathcal{L}$. As illustrated in Figure 2, CLARE requires no gradient computations or backward passes, making it faster and lightweight than existing approaches, enabling large-scale entanglement mapping.

While both CLARE and GradSim have a time complexity of $\mathcal{O}(L \cdot d^2)$, where d is the hidden dimension size and L is the number of MLP layers in an LLM, CLARE is substantially faster in practice as it avoids loss computation and backward passes. Moreover, GradSim requires $\mathcal{O}(L \cdot d^2)$ storage to retain full gradients for a fact, whereas CLARE only needs $\mathcal{O}(d)$, which is to store a single hidden state vector at layer $\mathcal{L}$. These compact representations allow CLARE to scale to thousands of facts efficiently. A detailed computational and storage complexity analysis is provided in Appendix A, and runtime cost analysis in Appendix B.

5 Experiments

Our curated corpus of 11,427 facts spans 212 unique prompt formats and 6,140 unique subjects,

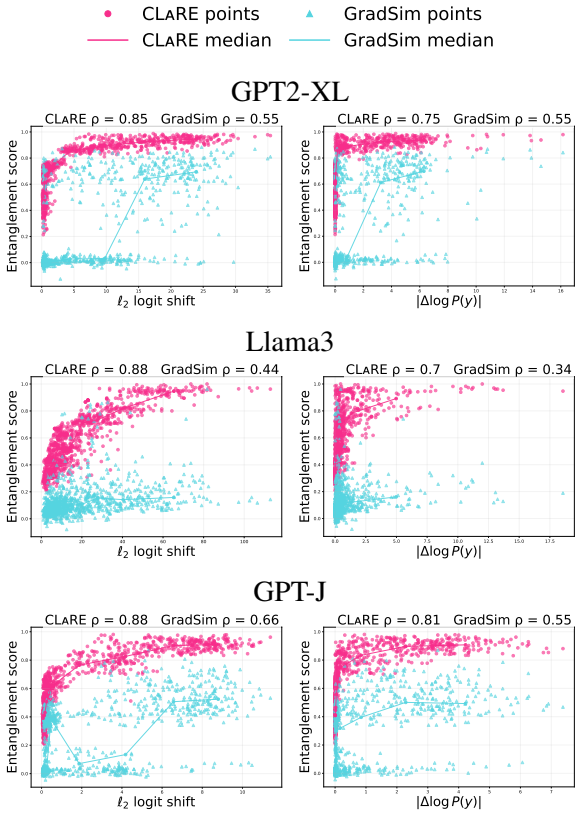


Figure 3: Correlation patterns for AlphaEdit: entanglement vs. ℓ_2 logit shift (left) and $|\Delta \log P(y)|$ (right).

enabling a systematic analysis of how ripple effects propagate globally across a model's knowledge base and how these dynamics vary across model architectures. All of our experiments are conducted on an NVIDIA H200 141GB GPU. We adopt the hyperparameter configurations outlined in the EasyEdit framework (Wang et al., 2024). We address three research questions here, and discuss more research questions in Appendix C.

RQ ①: How is CLARE's predictive accuracy and computational efficiency compared to gradient-based methods like GradSim?

RQ ②: How does the choice of layer affect CLARE's entanglement estimation quality?

RQ ③: How effectively can CLARE compute large-scale entanglement graphs to enable downstream tasks like budget-constrained red-teaming and building model-editing preservation sets?

5.1 Performance on entanglement estimation and computational efficiency (RQ ①)

To quantify the magnitude of ripple effects induced by a model edit, we measure how predictions on control facts—facts not directly edited—change before and after editing. Let θ and θ' be the parameters of the original and edited models, respectively. For control fact y with query q_y and original answer a_y, we define two ripple-effect magnitudes that capture both geometric and probabilistic shifts in model behavior:

(1) ℓ_2 logit shift captures the overall geometric shift in the LLM's output distribution for y:

$$\mathcal{L}_2(y) = \left\| f_\theta(q_y) - f_{\theta'}(q_y) \right\|_2 \tag{6}$$

where $f_\theta(q_y)$ and $f_{\theta'}(q_y)$ are the output logit vectors of the original and edited models for the input prompt q_y. A larger $\mathcal{L}_2(y)$ indicates stronger ripple effect on the model's internal representations.

(2) Original answer log-probability shift measures the probabilistic change of the answer a_y:

$$|\Delta \log P(y)| = \\ |\log P_{\theta'}(a_y \mid q_y) - \log P_\theta(a_y \mid q_y)| \tag{7}$$

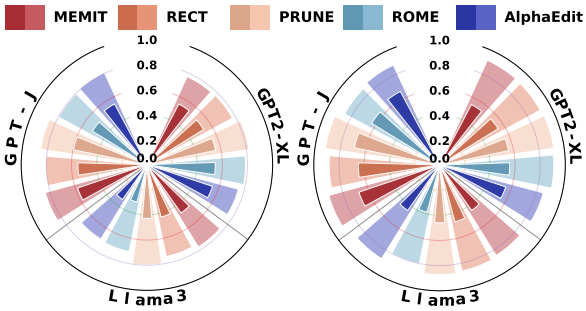
A larger $|\Delta \log P(y)|$ signifies that the LLM's belief about the control fact is altered (similar to Absolute Likelihood Gain in (Qin et al., 2024a)).

For each model, we sampled 1,000 edit–control pairs from our corpus, ensuring coverage across a broad spectrum of ℓ_2 logit shift magnitudes resulting from a single trial when edited by AlphaEdit (Fang et al., 2025). Each edit–control pair consists of an edit fact and a control fact. For each LLM and editing technique, we compute CLARE (Eq. 5) and GradSim entanglement scores for control facts with their corresponding edit facts, and compare them against both the ripple-effect magnitudes. We compute the Spearman rank correlation coefficient (ρ_s) between the entanglement scores and the observed ripple magnitudes. A higher ρ_s indicates that the method more accurately captures ripple effects. We evaluate five editing techniques (ROME (Meng et al., 2022), MEMIT (Meng et al., 2023), PRUNE (Ma et al., 2025), RECT (Gu et al., 2024), and AlphaEdit (Fang et al., 2025)) over three LLMs of varying sizes: GPT2-XL (1.5B), GPT-J (6B), and Llama3 (8B). Figure 3 shows these correlation patterns of these methods when edited by AlphaEdit. We present correlation patterns for other techniques and further analysis in Appendix D.

Table 1 reports ρ_s across all LLMs and editing techniques. While GradSim provides moderate correlation ($\rho_s \approx 0.45$–0.68), CLARE achieves consistently strong alignment ($\rho_s \approx 0.75$–0.92),

Table 1: Spearman Correlation Scores for CLARE and GradSim across $|\Delta \log P(y)|$ and ℓ_2 logit shift.

Model	Metric (ρ_s)	Method	Editing Techniques				
			MEMIT	RECT	ROME	PRUNE	AlphaEdit
GPT-J	$\lvert\Delta \log P(y)\rvert$	GradSim	0.582	0.549	0.509	0.597	0.547
		CLARE	**0.838** (+44.0%)	**0.805** (+46.6%)	**0.797** (+56.6%)	**0.854** (+43.0%)	**0.805** (+47.2%)
	ℓ_2 logit shift	GradSim	0.677	0.646	0.634	0.691	0.663
		CLARE	**0.913** (+34.9%)	**0.878** (+35.9%)	**0.857** (+35.2%)	**0.917** (+32.7%)	**0.875** (+32.0%)
GPT2-XL	$\lvert\Delta \log P(y)\rvert$	GradSim	0.545	0.531	0.544	0.558	0.550
		CLARE	**0.770** (+41.3%)	**0.774** (+45.8%)	**0.785** (+44.3%)	**0.822** (+47.3%)	**0.753** (+36.9%)
	ℓ_2 logit shift	GradSim	0.543	0.548	0.558	0.561	0.554
		CLARE	**0.915** (+68.5%)	**0.911** (+66.2%)	**0.909** (+62.9%)	**0.918** (+63.6%)	**0.854** (+54.2%)
Llama3	$\lvert\Delta \log P(y)\rvert$	GradSim	0.471	0.432	0.308	0.428	0.339
		CLARE	**0.772** (+63.9%)	**0.747** (+72.9%)	**0.706** (+129.2%)	**0.798** (+86.4%)	**0.697** (+105.6%)
	ℓ_2 logit shift	GradSim	0.442	0.468	0.389	0.460	0.442
		CLARE	**0.848** (+91.9%)	**0.861** (+84.0%)	**0.801** (+105.9%)	**0.871** (+89.3%)	**0.876** (+98.2%)



(a) ρ_s between entanglement values and $|\Delta \log P(y)|$.

(b) ρ_s between entanglement values and ℓ_2 logit shift.

Figure 4: Performance comparison between CLARE and GradSim for measuring editing entanglement. CLARE (wider, transparent bars) consistently achieves higher ρ_s than GradSim (narrower, solid bars).

with an average improvement of 62.2% over GradSim. For GPT2-XL, CLARE achieves an average improvement of 40.8% over GradSim. GPT-J exhibits similar strong predictive alignment, with mean gain of 53.1%. As shown in Figure 4, improvements are even more pronounced for Llama3, with CLARE achieving a mean 92.7% higher correlation than GradSim, indicating that CLARE generalizes well to even larger models.

To evaluate CLARE's computational cost, we track total processing time, peak GPU memory use, and the size of per-fact representations during entanglement computation. We benchmark these across seven LLMs: GPT2-XL (1.5B), Llama3.2 (3B), GPT-J (6B), Qwen2 (7B), Qwen2.5 (7B), Mistral (7B, v0.1), and Llama3 (8B). For all quantities, a lower value indicates better performance.

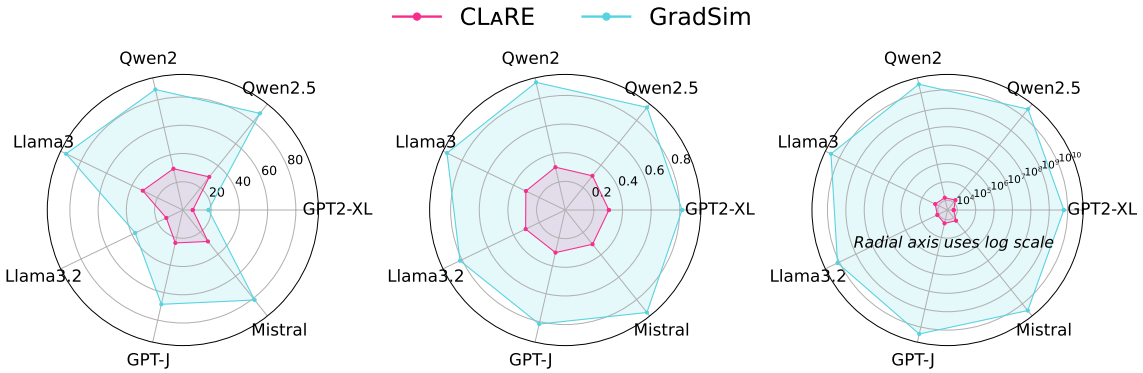
As shown in Table 2, CLARE achieves substantial speed improvements, achieving an average of **2.74×** speed-up over GradSim. CLARE's factual representations are extremely compact, being measured in kilobytes. On average, CLARE achieves over a **1.64-million-fold compression** relative to GradSim. Peak GPU memory use in CLARE is **2.85× lower** than GradSim. As observed in Figure 5, these results show that factual entanglement at the LLM's last critical layer reflects vulnerability to ripple propagation, and a lightweight, forward-only technique can outperform computationally intensive gradient methods.

5.2 Layer-wise Correlation Analysis (RQ ②)

To identify the most suitable layer to capture factual entanglement, we conduct a layerwise correlation study for the same five editing techniques, three models and edit-control pairs as in RQ ①. For each control fact, instead of only $\mathcal{L}$, we take each transformer layer l's hidden representation h_i^l, and find entanglement values using Eq. 5. For each l, we measure how well its scores predict ripple magnitude in control facts following edits, in terms of ℓ_2 logit shift. We compute the Spearman

Table 2: Performance comparison between CLARE and GradSim. Lower values for all quantities indicate better performance. The ratio rows (e.g., Speed-up) show how many times CLARE is faster or more memory-efficient than GradSim. For example, a speed-up of **3.0×** means CLARE completes the task 3.0 times faster than GradSim.

Method	GPT2-XL	Qwen2.5	Qwen2	Llama3	Llama3.2	GPT-J	Mistral
Total Processing Time (seconds)							
GradSim	0.796	0.915	0.868	0.917	0.774	0.763	0.833
CLARE	**0.305**	**0.305**	**0.307**	**0.305**	**0.307**	**0.305**	**0.306**
Speed-up	2.610×	3.000×	2.827×	3.007×	2.521×	2.502×	2.722×
Size of representation used for computation (bytes)							
GradSim	6.23e9	3.05e10	3.05e10	3.21e10	1.29e10	2.42e10	2.90e10
CLARE	**6400.00**	**14340.00**	**14340.00**	**16380.00**	**12290.00**	**16380.00**	**16380.00**
Compression	9.73e5×	2.13e6×	2.13e6×	1.96e6×	1.05e6×	1.48e6×	1.77e6×
Peak Utilised GPU Memory (GB)							
GradSim	18.212	87.513	87.513	91.765	37.445	68.474	81.483
CLARE	**7.016**	**29.981**	**29.981**	**31.539**	**12.995**	**23.905**	**28.478**
Reduction	2.596×	2.919×	2.919×	2.910×	2.881×	2.864×	2.861×



(a) Peak Utilised GPU Usage (GB) (b) Total Processing Time (seconds) (c) Size of representation (bytes)

Figure 5: Computational efficiency comparison. Closer to center => better performance.

correlation for each l's entanglement scores and ℓ_2 logit shift, yielding a per-layer correlation profile.

From figure 8, we observe a clear and consistent plateau around the last critical layer in correlation across all editing techniques, corresponding closely to the region previously identified to be storing factual associations by causal-tracing analyses (Meng et al., 2022, 2023). Quantitatively, Table 3 shows that for all three models and five editing techniques, the correlation computed at the final critical layer lies within **0–1 percentage points** of the global maximum observed across all layers. This indicates that ripple effects can most effectively be predicted when analysed using hidden representations of the deepest layer of the critical region, $\mathcal{L}$. We discuss $\mathcal{L}$-approximation in Appendix E in case if prior causal analysis of identifying critical layers is unavailable.

5.3 Performance on scalability and downstream applications support (RQ ③)

To demonstrate scalability, we release entanglement graphs for our entire corpus of 11,427 facts for GPT2-XL, GPT-J and Llama3. This is made possible by CLARE's compact vector representation, which occupies only a fraction of storage required by GradSim (see Table 2). These lightweight vectors are efficiently stored and quickly loaded to the GPU for pairwise cosine similarity computation for thousands of fact pairs. In contrast, GradSim requires storing the full gradient for each fact, which is equal in size to the model itself, making it infeasible for large-scale analysis. CLARE reduces this storage burden in orders of $10^6\times$, enabling corpus-wide analysis at scale. We analyse distribution of the corpus's facts and their entanglement patterns in Appendix C.2.

To support comprehensive evaluation of editing techniques under a constrained evaluation budget, we identify highly entangled facts in our cor-

Table 3: ρ_s between CLARE and ripple effect (ℓ_2 logit shift) in control facts observed at $\mathcal{L}$, compared to maximum value possible at any layer. "Difference from peak" reports the absolute deviation in percentage points (p.p.).

Model	Layer Type	AlphaEdit	MEMIT	PRUNE	RECT	ROME
Llama3	$\mathcal{L} = 9$	65.16%	63.76%	64.35%	61.73%	70.54%
	Maximum ρ_s (any layer)	65.16%	64.16%	64.84%	61.73%	70.54%
	Difference from peak (p.p.)	0.00	0.40	0.49	0.00	0.00
GPT-J	$\mathcal{L} = 9$	89.30%	80.13%	92.72%	90.26%	88.46%
	Maximum ρ_s (any layer)	89.72%	80.13%	92.97%	90.26%	88.58%
	Difference from peak (p.p.)	0.42	0.00	0.25	0.00	0.12
GPT2-XL	$\mathcal{L} = 18$	76.61%	87.61%	87.98%	86.33%	85.10%
	Maximum ρ_s (any layer)	77.49%	87.67%	87.99%	86.42%	85.26%
	Difference from peak (p.p.)	0.88	0.06	0.01	0.09	0.16

1. The name of the country of citizenship of **Kate Winslet** is UK (affects 1257 facts)
2. The place of burial of **Audrey Hepburn** is Cemetery of Tolochenaz (affects 1233 facts)
3. The place of birth of **Audrey Hepburn** is Rue Keyenveld (affects 1222 facts)
4. The place of death of **Audrey Hepburn** is Tolochenaz (affects 1218 facts)
5. The name of the country of citizenship of **Mark Wahlberg** is USA (affects 1204 facts)

Figure 6: Top-5 entangled facts in GPT-J in our corpus.

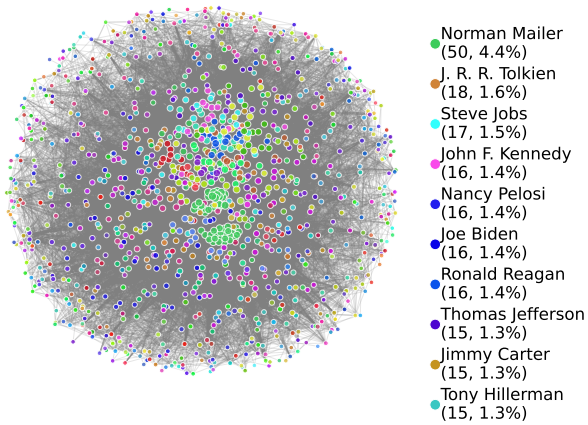


Figure 7: An entanglement cluster in GPT-J with its ten most entangled subjects listed with their number and percentage of facts. It consists of 1,149 facts and 397 subjects, with 87.9% of edges between inter-subject facts. Node color indicates subject identity.

pus. These high-density facts are critical pressure points in LLMs, as editing them may trigger widespread ripple effects. Figure 6 shows the top 5 most entangled facts in GPT-J, ranked by the number of other facts they are connected to with an entanglement score above 0.7. As shown in Appendix C.1, facts exceeding this threshold are highly ripple-prone, serving as ideal adversarial candidates. Thus, by surfacing these highly connected regions, CLARE enables fact prioritization for red-teaming. We present more clusters for various models in Appendix F.

We cluster the entanglement graphs using the Louvain community detection algorithm, where nodes represent facts and edges connect pairs with high entanglement scores (cosine similarity > 0.7). Figure 7 shows one such cluster in GPT-J, highlighting the top 10 subjects whose facts have the highest connectivity and thus the strongest ripple potential. These clusters span multiple subjects, revealing substantial cross-entity interference. They are especially useful for constructing preservation sets: when editing a fact about subject A, constraints must be applied to other facts within the same cluster to minimize collateral degradation. This extends the capabilities of DiKE (Zhang et al., 2025), enabling it to preserve facts in hidden space as well. We present more clusters for various LLMs in Appendix G.

6 Conclusion

We presented CLARE, a lightweight and scalable technique to estimate factual entanglement within LLMs. CLARE uses forward activations from a single intermediate layer, without requiring model gradients. Through extensive model-editing experiments, we show that CLARE achieves stronger alignment with observed ripple effects than gradient-based baselines, with an average 62.2% improvement in Spearman correlation, while offering 2.74× speedups, 2.85× lower peak GPU memory usage, and requires only a fraction of the storage needed to preserve fact representations. CLARE also enables downstream tasks such as budget-aware red-teaming and preservation set generation. Together, these contributions advance the understanding of ripple-effects in LLMs and support CLARE as a practical foundation for responsible model editing.

7 Limitations

While CLARE provides a scalable and empirically strong proxy for predicting ripple effects through activation-based entanglement scores, our analysis remains correlational in nature. That is, although we observe high alignment between CLARE scores and observed behavioral interference post-editing, we do not establish a formal causal mechanism linking representational entanglement to model generalization or degradation. Investigating the causal pathways by which knowledge is stored, shared, and disrupted during pretraining and editing remains an open and important direction for future work. A second limitation is that CLARE currently functions as a diagnostic tool rather than an editing strategy. While CLARE enables the construction of entanglement-aware preservation sets, we do not integrate these constraints into existing editing techniques. Future research could build on CLARE's preservation sets to create entanglement-aware editing techniques that reduce collateral degradation during updates.

8 Ethical considerations

From an ethical standpoint, the ability to map these regions allows developers to be more accountable for the "collateral degradation" caused by their updates. By building preservation sets based on entanglement clusters, researchers can actively prevent the erosion of model integrity. However, the paper acknowledges that CLARE is currently a diagnostic tool, and the ethical responsibility remains with the practitioners to integrate these constraints into the actual editing process to ensure a "preventive" rather than "reactive" approach to model safety. We emphasize that our release of entanglement graphs is intended for research and safety applications only, and we encourage future work to adopt responsible deployment practices. Incorporating entanglement-aware auditing may play a key role in developing accountable model-editing pipelines in high-stakes or sensitive domains. All entanglement graphs and data samples are released under the MIT License.

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Mindful-rag: A study of points of failure in retrieval augmented generation. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 607–611.

Manit Baser, Dinil Mon Divakaran, and Mohan Gurusamy. 2025. Thinkeval: Practical evaluation of knowledge preservation and consistency in llm editing with thought-based knowledge graphs. *arXiv preprint arXiv:2506.01386*.

Qizhou Chen, Taolin Zhang, Xiaofeng He, Dongyang Li, Chengyu Wang, Longtao Huang, and Hui Xue'. 2024. Lifelong knowledge editing for LLMs with retrieval-augmented continuous prompt learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13565–13580, Miami, Florida, USA. Association for Computational Linguistics.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2025. Everything is editable: Extend knowledge editing to unstructured data in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR 2025)*.

Zilu Dong, Xiangqing Shen, Zinong Yang, and Rui Xia. 2025. ChainEdit: Propagating ripple effects in LLM knowledge editing through logical rule-guided chains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13558–13571, Vienna, Austria. Association for Computational Linguistics.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained knowledge editing for language models. In *Proceedings of the International Conference on Learning Representations, 2025*. https://openreview.net/attachment?id=HvSytvg3Jh&name=pdf, accessed 2025-12-06.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16801–16819, Miami, Florida, USA. Association for Computational Linguistics.

Nicolas Guerin, Ryan M. Nefdt, and Emmanuel Chemla. 2025. Qualifying knowledge and knowledge sharing in multilingual models. *Transactions on Machine Learning Research*.

Akshat Gupta, Dev Sajnani, and Gopala Anumanchipalli. 2024. A unified framework for model editing. *arXiv*, abs/2403.14236.

Akshat Gupta, Nehal Sindhu, and Gopala Anumanchipalli. 2025. How linearly associative are memories in large language models? In *New Frontiers in Associative Memories*.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36:47934–47959.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. 2023. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11548–11559, Toronto, Canada. Association for Computational Linguistics.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-patcher: One mistake worth one neuron. In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2025. Investigating bias in LLM-based bias detection: Disparities between LLMs and human perception. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10634–10649, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiateng Liu, Pengfei Yu, Yuji Zhang, Sha Li, Zixuan Zhang, Ruhi Sarikaya, Kevin Small, and Heng Ji. 2024. EVEDIT: Event-based knowledge editing for deterministic knowledge propagation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4907–4926, Miami, Florida, USA. Association for Computational Linguistics.

Jiaxiang Liu, Boxuan Xing, Chenhao Yuan, ChenxiangZhang ChenxiangZhang, Di Wu, Xiusheng Huang, Haida Yu, Chuhan Lang, Pengfei Cao, Jun Zhao, and Kang Liu. 2025. Know-MRI: A knowledge mechanisms revealer&interpreter for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 199–210, Vienna, Austria. Association for Computational Linguistics.

Jun-Yu Ma, Hong Wang, Hao-Xiang Xu, Zhen-Hua Ling, and Jia-Chen Gu. 2025. Perturbation-restrained sequential model editing. In *Proceedings of the International Conference on Learning Representations (ICLR 2025)*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *Proceedings of the International Conference on Learning Representations, 2023*. Available at url-https://iclr.cc/virtual/2023/oral/12726.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *Proceedings of the International Conference on Learning Representations (ICLR 2022)*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.

Jiaxin Qin, Zixuan Zhang, Chi Han, Manling Li, Pengfei Yu, and Heng Ji. 2024a. Why does new knowledge create messy ripple effects in LLMs? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*.

Jiaxin Qin, Zixuan Zhang, Chi Han, Pengfei Yu, Manling Li, and Heng Ji. 2024b. Why does new knowledge create messy ripple effects in LLMs? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12602–12609, Miami, Florida, USA. Association for Computational Linguistics.

Mubashar Raza, Zarmina Jahangir, Muhammad Bilal Riaz, Muhammad Jasim Saeed, and Muhammad Awais Sattar. 2025. Industrial applications of large language models. *Scientific Reports*, 15(1):13755.

Domenic Rosati, Robie Gonzales, Jinkun Chen, Xuemin Yu, Yahya Kayani, Frank Rudzicz, and Hassan Sajjad. 2024. Long-form evaluation of model editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3749–3780, Mexico City, Mexico. Association for Computational Linguistics.

Dorsaf Sallami, Yuan-Chen Chang, and Esma Aïmeur. 2024. From deception to detection: The dual roles of large language models in fake news. *arXiv preprint arXiv:2409.17416*.

Jianchen Wang, Zhouhong Gu, Xiaoxuan Zhu, Lin Zhang, Haoning Ye, Zhuozhi Xiong, Sihang Jiang, Hongwei Feng, and Yanghua Xiao. 2025. The missing piece in model editing: A deep dive into the hidden damage brought by model editing. In

ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2024. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–93, Bangkok, Thailand. Association for Computational Linguistics.

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. 2023. Foundation models for decision making: Problems, methods, and opportunities. In *NeurIPS 2023 Workshop on Foundation Models for Decision Making*. ArXiv:2303.04129.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Defne Yigci, Merve Eryilmaz, Ail K. Yetisen, Savas Tasoglu, and Aydogan Ozcan. 2025. Large language model-based chatbots in higher education. *Advanced Intelligent Systems*, 7(3):2400429.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, and Zhumin Chen. 2024a. Knowledge graph enhanced large language model editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22647–22662, Miami, Florida, USA. Association for Computational Linguistics.

Mengqi Zhang, Zisheng Zhou, Xiaotian Ye, Qiang Liu, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2025. Disentangling knowledge representations for large language model editing. *arXiv preprint arXiv:2505.18774*.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024b. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zihao Zhao, Yuchen Yang, Yijiang Li, and Yinzhi Cao. 2024. RippleCOT: Amplifying ripple effect of knowledge editing in language models via chain-of-thought in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6337–6347, Miami, Florida, USA. Association for Computational Linguistics.

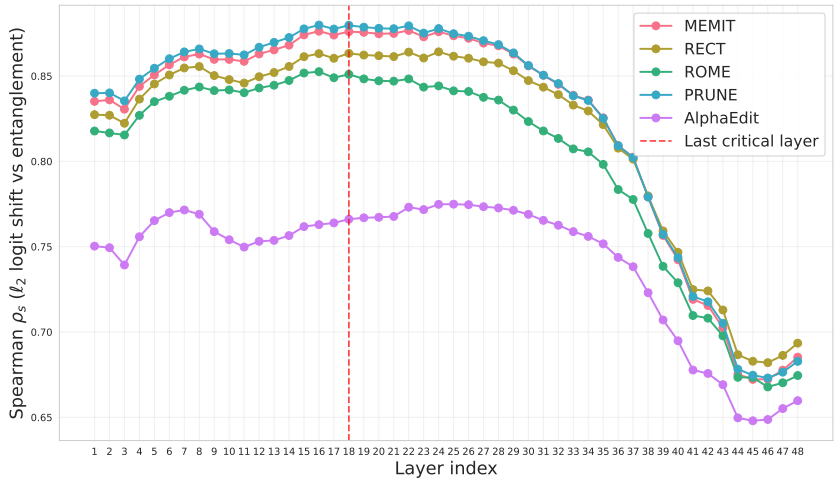
Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.

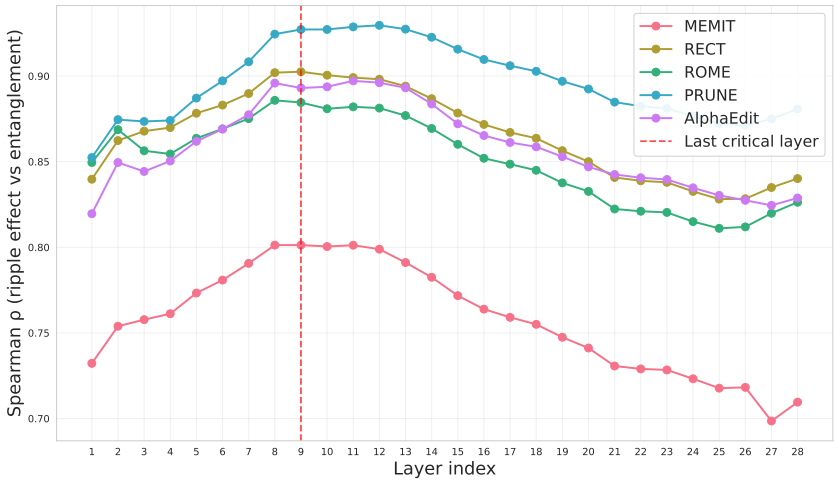
A Computational complexity analysis

We analyze the time and space complexity of GradSim and CLARE to compute entanglement between two facts. For GradSim, the process can be divided into primarily into four components: forward pass, loss computation, backward pass (for both facts), and then cosine similarity computation over the parameter gradients. CLARE requires only a forward pass up to the last critical layer $\mathcal{L}$ for both facts, followed by cosine similarity computation for the hidden representation at $\mathcal{L}$. We provide the details for each step in GradSim and CLARE in Table 4. GradSim requires both a forward and backward pass per factual triplet, while CLARE achieves comparable entanglement estimation using only a single forward pass.

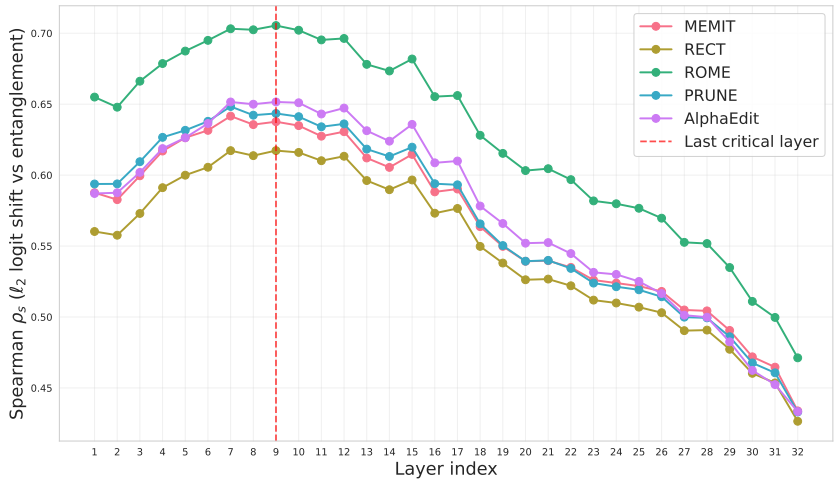
Time Complexity. For GradSim, the forward pass processes N input tokens through L transformer layers, with each layer usually having the same hidden dimension size d. Each layer consists of multi-head attention with complexity $\mathcal{O}(N^2 \cdot d)$ and feed-forward MLPs with complexity $\mathcal{O}(d^2)$, yielding a total forward pass complexity of $\mathcal{O}(L \cdot d^2 + N^2 \cdot L \cdot d)$. For typical model-editing prompts where $N \approx$ 10–20 tokens and d is in the orders of 10^3, (4096 in the case of Llama3-8B), the MLP term dominates since $d^2 \gg N^2 \cdot d$. Thus, the forward pass simplifies to $\mathcal{O}(L \cdot d^2)$. The loss computation involves a final projection from hidden dimension d to vocabulary size v, requiring $\mathcal{O}(d \cdot v)$ operations. This is also typically negligible compared to the $\mathcal{O}(L \cdot d^2)$ forward pass cost. The backward pass has the same computational complexity as the forward pass, $\mathcal{O}(L \cdot d^2)$, as it traverses in reverse to compute gradients with respect to all parameters. Finally, computing the cosine similarity between gradient vectors requires $\mathcal{O}(L \cdot d^2)$ oper-



(a) GPT2-XL



(b) GPT-J



(c) Llama3

Figure 8: Layerwise Spearman correlation (ρ_s) between CLARE and observed ripple magnitudes. Correlation peaks around the last critical layer indicate that it is most informative about entanglement estimation.

Table 4: Step-by-step comparison of time and space complexity between GradSim and CLARE for computing entanglement between two facts.

Component	GradSim	CLARE
Forward pass (per fact)	$\mathcal{O}(L \cdot d^2 + N^2 \cdot L \cdot d)$	$\mathcal{O}(L \cdot d^2 + N^2 \cdot L \cdot d)$
Loss computation (per fact)	$\mathcal{O}(d \cdot v)$	—
Backward pass (per fact)	$\mathcal{O}(L \cdot d^2)$	—
Cosine similarity computation	$\mathcal{O}(L \cdot d^2)$	$\mathcal{O}(d)$
Total Time Complexity	$\sim \mathcal{O}(L \cdot d^2)$	$\sim \mathcal{O}(L \cdot d^2)$
Gradient storage (per fact)	$\mathcal{O}(L \cdot d^2)$	—
Activation storage (per fact)	$\mathcal{O}(N \cdot L \cdot d)$	$\mathcal{O}(d)$
Total Space Complexity (per fact)	$\sim \mathcal{O}(L \cdot d^2)$	$\mathcal{O}(d)$

ations since we must compute the dot product and norms over all $\mathcal{O}(L \cdot d^2)$ model parameters. The total time complexity per fact pair is thus $\mathcal{O}(L{\cdot}d^2)$, dominated by the forward and backward passes.

For CLARE, the forward pass up to layer $\mathcal{L}$ has complexity $\mathcal{O}(L \cdot d^2 + N^2 \cdot L \cdot d) \approx \mathcal{O}(L \cdot d^2)$. No backward pass is required, as CLARE operates purely on forward activations. The cosine similarity computation operates on d-dimensional hidden state vectors (one per fact), requiring only $\mathcal{O}(d)$ operations to compute the dot product $\langle h_i^{\mathcal{L}}, h_j^{\mathcal{L}} \rangle$ and vector norms $\|h_i^{\mathcal{L}}\|$ and $\|h_j^{\mathcal{L}}\|$. While CLARE and GradSim have the same time complexity, CLARE is ∼2.8× faster as it eliminates the need for backward passes.

Space Complexity. GradSim requires $\mathcal{O}(L \cdot d^2)$ space to store gradients for all model parameters (weight matrices of dimension $d \times d$ across L layers), plus $\mathcal{O}(N \cdot L \cdot d)$ space for caching intermediate activations needed during backpropagation. In practice, gradient storage dominates, yielding $\mathcal{O}(L \cdot d^2)$ total space complexity. In contrast, CLARE requires only $\mathcal{O}(d)$ space to store a single d-dimensional hidden state vector per fact at layer $\mathcal{L}$. This efficiency enables CLARE to scale to analyzing thousands of fact pairs, as each fact requires only a compact d-dimensional representation rather than the full gradient footprint, and thus CLARE achieves an average of 1.64-million-fold compression in storing fact representations.

B Runtime cost analysis

We empirically benchmark the component-wise computational costs of CLARE and GradSim across seven language models. We measure both execution time and peak GPU memory usage for each step involved in computing entanglement between a pair of facts.

Runtime Analysis. Table 5 reports the time (in seconds) required by each step in CLARE and GradSim. CLARE involves two forward passes (one for each fact) followed by cosine similarity computation over d-dimensional activation vectors. GradSim, in contrast, requires forward, loss, and backward passes for both facts, followed by cosine similarity computation over full gradients. Across all models, CLARE achieves substantial speedups. This efficiency gain stems primarily from the elimination of loss calculation and backward passes.

Memory Utilization. Table 6 presents the peak GPU memory usage (in GB) at each stage of computation. For CLARE, peak memory remains below 32 GB even for 7B and 8B models, enabling full execution on a single H200 GPU. In contrast, GradSim exhibits significantly higher memory demands, peaking near or beyond 90 GB for large models due to the need to retain gradients and activations for all parameters.

As shown in Table 2, CLARE stores representations of size $4 \times d$, where d is the hidden dimension of the LLM. In contrast, GradSim stores the full gradient for each fact, with a storage cost of $4 \times P$, where P is the total number of model parameters. All storage measurements are computed using FP32 precision.

C More research questions

We address the following research questions here:

RQ ④: Can CLARE serve as a diagnostic for evaluating the disruptive impact of model-editing techniques and comparing them?

Table 5: Detailed time breakdown (seconds) for CLARE and GradSim. CLARE follows: Forward Pass ($\times 2$) $\oplus$ Cosine similarity (Sim.) of two vectors (Vec). GradSim follows: Forward ($\times 2$) $\oplus$ Loss ($\times 2$) $\oplus$ Backward ($\times 2$) $\oplus$ Cosine similarity of gradients (Grad).

Model	**Method 1: CLARE**			**Method 2: GradSim**				
	Forward ($\times 2$ Vec)	Sim.	**Total**	Forward ($\times 2$ Grad)	Loss ($\times 2$ Grad)	Backward ($\times 2$ Grad)	Sim.	**Total**
GPT-J	0.204	0.101	**0.305**	0.251	0.205	0.205	0.102	**0.763**
GPT2-XL	0.204	0.101	**0.305**	0.284	0.205	0.205	0.102	**0.796**
Llama3.2	0.206	0.101	**0.307**	0.262	0.204	0.205	0.102	**0.773**
Llama3	0.204	0.101	**0.305**	0.305	0.205	0.205	0.202	**0.917**
Mistral	0.205	0.101	**0.306**	0.304	0.205	0.223	0.101	**0.833**
Qwen2	0.206	0.101	**0.307**	0.255	0.204	0.206	0.202	**0.867**
Qwen2.5	0.204	0.101	**0.305**	0.303	0.205	0.205	0.202	**0.915**

RQ ⑤: How are the entanglement patterns across different language models?

C.1 Evaluating the Disruptive Impact of Model-Editing Techniques (RQ④)

Using CLARE, we qualitatively compare editing techniques by their propensity to induce ripple effects. Appendix D shows CLARE's entanglement correlation patterns for a single edited fact and many control facts, offering a finer-grained view of ripple susceptibility. They show a consistent threshold across models and techniques: entanglement scores below 0.7 yield minimal change, while scores more than 0.7 lead to sharply increasing ripple effects. This nonlinear transition shows that highly entangled facts are more prone to ripple effects, a risk CLARE can detect prior to editing. When comparing editing techniques across GPT2-XL, GPT-J and Llama3, we observe clear differences in ripple severity. ROME and PRUNE often induce stronger and more widespread ripple effects as control facts become more entangled, whereas RECT and MEMIT generally yield more localized, controlled updates. AlphaEdit exhibits intermediate behavior. Thus, CLARE both anticipates where ripple effects will occur and provides a model- and technique-agnostic way to assess robustness and disruptiveness in editing, supporting informed deployment and design decisions.

C.2 Entanglement score Distribution (RQ ⑤)

Figure 9 shows cosine similarity distribution between CLARE representations across all 11,427 factual triplets for GPT-J, GPT2-XL and Llama3. Most entanglement scores lie between 0.2 and 0.6, indicating moderate representational overlap between most fact pairs. Notably, a non-trivial mass exists near the upper extreme (0.95-1.0), suggesting a high-density zone of strongly entangled facts that may be especially susceptible to ripple effects. GPT2-XL also shows a distinct secondary cluster with negative cosine similarities (centered near -0.35), not present in the other models, pointing to architectural differences in how subject representations are organized. The long-tailed shape and multimodal structure of the distribution underscore the need for selective entanglement-aware editing strategies.

D Performance of other editing techniques in RQ ①

Figures 14, 15, 16, 17, 18, 19, 20 and 21 show the correlation patterns of GradSim and CLARE for different models when edited by different edited techniques over the 1,000 samples as discussed in RQ ①. Figures 22, 23, 24, 25, 26, 27, 28, 29, 30, and 31 illustrate CLARE's entanglement correlation patterns using a single edited fact probed against multiple control facts, to capture a more fine-grained view of ripple susceptibility. Across all editing techniques, models and metrics (ℓ_2 logit shift and $|\Delta \log P(y)|$), we observe a consistent increasing trend: higher CLARE entanglement scores correspond to larger ripple effects. This monotonic relationship supports the reliability of CLARE as a forward-only proxy for ripple-prone regions, generalizing robustly across diverse settings without the need for gradient computation. Across all settings, we observe a thresholded ris-

Table 6: Step-by-step peak GPU memory consumption (GB). CLARE steps: Fwd (forward pass) 1 → Fwd 2 → Sim (Cosine similarity of two vectors). **GradSim** steps: Gradient 1 (Fwd → Loss calculation → Bwd (backward pass) → Gradient 2 (Fwd → Loss → Bwd) → Sim (Cosine similarity of two gradients).

Model	**Method 1: CLARE**			**Method 2: GradSim**						
	Forward Pass		Sim	Gradient 1 Steps			Gradient 2 Steps			Sim
	Fwd 1	Fwd 2		Fwd	Loss	Bwd	Fwd	Loss	Bwd	
GPT-J	23.815	**23.905**	22.588	33.551	33.550	45.167	56.134	56.133	67.710	**68.474**
GPT2-XL	6.968	**7.016**	6.013	9.032	9.029	12.275	15.003	15.000	18.210	**18.212**
Llama3.2	12.940	**12.995**	12.001	18.042	18.037	25.475	30.048	30.043	37.439	**37.445**
Llama3	31.459	**31.539**	29.946	44.012	44.007	59.898	73.967	73.961	89.813	**91.765**
Mistral	28.404	**28.478**	27.009	40.331	40.330	54.021	67.348	67.346	80.999	**81.483**
Qwen2	29.882	**29.981**	28.541	41.820	41.815	57.023	70.316	70.310	85.491	**87.513**
Qwen2.5	29.882	**29.981**	28.541	41.820	41.815	57.023	70.316	70.310	85.491	**87.513**

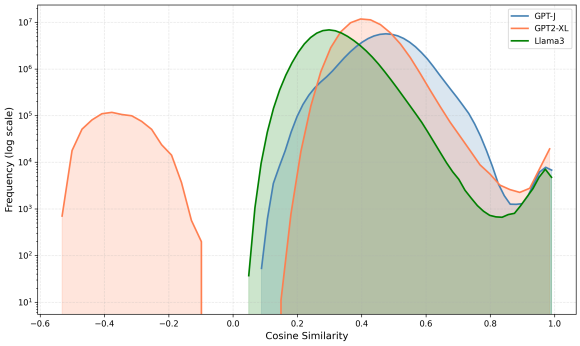


Figure 9: Distribution of cosine similarities between CLARE representations across 11,427 facts.

ing pattern. Entanglement scores below approximately 0.7 correspond to negligible or minimal changes in behavior (flat region), whereas scores above this threshold consistently lead to sharply increasing ripple effects. This pattern consistency across models, techniques, and evaluation metrics assists in separating safe vs. high-risk edits based purely on forward activations.

We hypothesize that the observed threshold around 0.7 cosine similarity reflects a critical geometric boundary in the model's representational manifold. In high-dimensional space, a cosine similarity of 0.7 corresponds to an angular separation of approximately $45°$, suggesting that facts within this cone may share overlapping subspaces or "knowledge circuits" (Yao et al., 2024). Edits that shift one fact's representation may inadvertently affect other facts within this angular neighborhood, manifesting as ripple effects. Further investigation is needed to validate whether this threshold consistently emerges across different model architectures.

E Approximation of the last critical layer

When causal-tracing data or prior analyses identifying the factual-memory layers are unavailable, we approximate the final critical layer using a simple depth-based heuristic. Prior work and our replication of the EasyEdit repository indicate that critical layers across a range of transformer architectures typically occur within the lower-to-middle portion of the model—specifically, around one-third of the total depth. As shown in Table 8, the ratio of the last critical layer to total layers spans approximately **0.28–0.38** across diverse architectures, including GPT-J, GPT-2-XL, Llama, Qwen, Mistral, and Baichuan. This empirical regularity supports the use of the one-third-depth index as a reliable proxy for the last causal layer when such information is unknown.

Thus, we take the layer at one-third of the model's total depth as the *approximated critical layer* for computing CLARE scores. This approximation enables scalable entanglement estimation without requiring causal-tracing interventions or proprietary layer-level annotations. Table 7 reports the Spearman correlation (ρ_s) between CLARE scores and empirically measured ripple magnitudes when CLARE is computed from this one-third-depth layer. Across all three LLMs (GPT2-XL, GPTJ-6B, and Llama3-8B), the correlation at the approximated layer remains within approximately **0–1.5 percentage points of the global peak**, indicating that the heuristic provides a robust and practical substitute for exact causal-layer identification. Notably, GPT-J and GPT-2-XL exhibit almost no degradation relative to their optimal layer correlations (differences $<$

Model	Layer Type	AlphaEdit	MEMIT	PRUNE	RECT	ROME
Llama-3	1/3-depth layer = 11	64.31%	62.74%	63.41%	61.02%	69.53%
	Maximum ρ_s (any layer)	65.16%	64.16%	64.84%	61.73%	70.54%
	Difference from peak (p.p.)	0.85	1.42	1.43	0.71	1.01
GPT-J	1/3-depth layer = 9	89.30%	80.13%	92.72%	90.26%	88.46%
	Maximum ρ_s (any layer)	89.72%	80.13%	92.97%	90.26%	88.58%
	Difference from peak (p.p.)	0.42	0.00	0.25	0.00	0.12
GPT2-XL	1/3-depth layer = 16	76.29%	87.63%	87.99%	86.32%	85.26%
	Maximum ρ_s (any layer)	77.49%	87.67%	87.99%	86.42%	85.26%
	Difference from peak (p.p.)	1.20	0.04	0.00	0.10	0.00

Table 7: Spearman correlation (ρ_s) between CLARE and ripple effect (ℓ_2 logit shift) in control facts at the one-third-depth layer (indexed by layer number), compared with the maximum value observed across all layers for each editing method. "Difference from peak" reports the absolute deviation in percentage points (p.p.).

Model	Index of $\mathcal{L}$ / Number of layers
Baichuan-7B	0.28125
GPT-J-6B	0.32143
GPT-2-XL	0.37500
InternLM-7B	0.28125
Llama-7B	0.28125
Llama3.2-3B	0.34615
Mistral-7B	0.28125
Qwen-7B	0.28125
Qwen2.5-7B	0.32143

Table 8: Ratio of the last critical layer to total number of layers for various models, taken from the EasyEdit repository. Values cluster around one-third depth, supporting the use of a one-third-depth heuristic when the critical layer is unknown.

0.5 p.p. for most methods), while Llama-3-8B shows slightly higher but still marginal deviations (1.0–1.4 p.p.). This suggests that representational entanglement is distributed smoothly around the critical band, and that the one-third-depth heuristic yields near-optimal CLARE estimates.

F Cost-effective red-teaming

To systematically identify areas of elevated editing risk, we analyse our corpora to find facts with exceptionally high representational connectivity, i.e., those whose internal representations overlap with a large number of other facts. These high-density knowledge nodes form critical pressure points within the model. Editing them is likely to induce observable widespread ripple effects. Figures 10, 11, 12 and 13 list the 50 most entangled facts in GPT2-XL, Llama3 and GPT-J, showing the number of connections each fact has within the graph with entanglement score more than 0.7. Such facts are not only more vulnerable to unintended interference, but also ideal candidates for red-teaming, or inclusion in post-edit evaluation suites. By spotlighting these highly connected regions, CLARE supports the prioritization of facts for testing under constrained evaluation budgets.

G Entanglement clusters

To analyse large-scale entanglement clusters, we construct graphs where nodes represent factual triplets and edges connect pairs with high entanglement scores (cosine similarity > 0.7). As observed in Appendix D, they have high tendency of ripple susceptibility. We apply the Louvain community detection algorithm to our corpora, identifying densely connected groups of entanglement clusters of triplets. Figures 32a, 32b, 33a, 33b, 33c, 33d, 33e, 33f, 34a, 34b, 34c, 34d, 34e, 35a, 35b, 35c, 35d, 35e, 35f, 36, 37a, 37b, 37c, 37d, 38a, 38b and 38c illustrate these clusters for GPT-J, GPT2-XL and Llama3. All of these contain 10 subjects that have strongest ripple potential because changing that fact influences the largest number of other facts. These clusters reflect internal representational overlap, and thus pinpoint regions where edits to one fact are likely to induce widespread ripple effects. We perform this clustering for Llama3, GPT2-XL, and GPT-J, and visualize all clusters containing more than 50 facts. Nodes are colored by subject, allowing visual identification of dominant subjects and high-risk hubs. This analysis reveals latent knowledge bottlenecks where facts are densely entangled, guiding safer and more targeted editing.

1. The name of the head of government of Spain is -> Pedro Sánchez → affects 296 other facts
2. The name of the head of government of France is -> Élisabeth Borne → affects 289 other facts
3. The name of the head of state of France is -> Emmanuel Macron → affects 273 other facts
4. The name of the current head of state in France is -> Emmanuel Macron → affects 271 other facts
5. The name of the head of state of Spain is -> Felipe VI of Spain → affects 266 other facts
6. The name of the capital city of France is -> Paris → affects 258 other facts
7. The name of the anthem of France is -> La Marseillaise → affects 252 other facts
8. The name of the capital city of Spain is -> Madrid → affects 251 other facts
9. The official language of Italy is -> Italian → affects 250 other facts
10. The name of the currency in Poland is -> Złoty → affects 250 other facts
11. The official language of Spain is -> Spanish → affects 247 other facts
12. The official language of France is -> French → affects 243 other facts
13. The name of the head of government of Poland is -> Mateusz Morawiecki → affects 242 other facts
14. The official language of Poland is -> Polish → affects 242 other facts
15. The official language of Germany is -> German → affects 240 other facts
16. The name of the capital city of Russia is -> Moscow → affects 239 other facts
17. The name of the currency in Sri Lanka is -> Sri Lankan rupee → affects 238 other facts
18. The name of the head of state of Poland is -> Andrzej Duda → affects 236 other facts
19. The name of the currency in Spain is -> euro → affects 233 other facts
20. The official language of Sri Lanka is -> Sinhala → affects 230 other facts
21. The name of the head of government of India is -> Narendra Modi → affects 230 other facts
22. The name of the head of state of Sri Lanka is -> Ranil Wickremesinghe → affects 229 other facts
23. The name of the capital city of occupation of Japan is -> Tokyo → affects 224 other facts
24. The name of the capital city of Poland is -> Warsaw → affects 224 other facts
25. The name of the head of government of Sri Lanka is -> Ranil Wickremesinghe → affects 222 other facts
26. The official language of Australia is -> English → affects 220 other facts
27. The official language of Slovakia is -> Slovak → affects 216 other facts
28. The official language of Romania is -> Romanian → affects 215 other facts
29. The name of the current head of state in Portugal is -> Marcelo Rebelo de Sousa → affects 215 other facts
30. The name of the head of state of Slovakia is -> Zuzana Čaputová → affects 214 other facts
31. The name of the anthem of Sri Lanka is -> Sri Lanka Matha → affects 213 other facts
32. The capital of Lithuania is -> Vilnius → affects 210 other facts
33. The capital of Indonesia is -> Jakarta → affects 208 other facts
34. The name of the currency in India is -> Indian rupee → affects 207 other facts
35. The name of the currency in Slovakia is -> euro → affects 206 other facts
36. The official language of Argentina is -> Spanish → affects 206 other facts
37. The name of the head of government of Slovakia is -> Eduard Heger → affects 205 other facts
38. The name of the head of state of India is -> Droupadi Murmu → affects 205 other facts
39. The capital of Romania is -> Bucharest → affects 205 other facts
40. The name of the head of state of Slovenia is -> Nataša Pirc Musar → affects 204 other facts
41. The official language of India is -> Hindi → affects 204 other facts
42. The name of the currency in Ukraine is -> Hryvnia → affects 203 other facts
43. The name of the head of government of Slovenia is -> Robert Golob → affects 202 other facts
44. Louis XVII of France died in the city of -> Paris → affects 202 other facts
45. The name of the anthem of Slovakia is -> Nad Tatrou sa blýska → affects 201 other facts
46. The official language of Peru is -> Spanish → affects 201 other facts
47. The official language of Slovenia is -> Slovene → affects 201 other facts
48. Louis XV of France is affiliated with the religion of -> Catholic Church → affects 199 other facts
49. The official language of Japan is -> Japanese → affects 198 other facts
50. The name of the capital city of Slovakia is -> Bratislava → affects 196 other facts

Figure 10: Most entangled facts ranked by representational connectivity in GPT2-XL.

1. The name of the head of government of Poland is -> Mateusz Morawiecki → affects 89 other facts
2. The official language of Poland is -> Polish → affects 85 other facts
3. The official language of Romania is -> Romanian → affects 78 other facts
4. The name of the head of state of Poland is -> Andrzej Duda → affects 77 other facts
5. The official language of Germany is -> German → affects 73 other facts
6. The name of the capital city of Poland is -> Warsaw → affects 71 other facts
7. The official language of Italy is -> Italian → affects 68 other facts
8. The official language of Ukraine is -> Ukrainian → affects 67 other facts
9. The name of the head of government of Spain is -> Pedro Sánchez → affects 66 other facts
10. The name of the head of government of Turkey is -> Recep Tayyip Erdoğan → affects 66 other facts
11. The official language of Spain is -> Spanish → affects 63 other facts
12. The name of the capital city of Ukraine is -> Kyiv → affects 62 other facts
13. The name of the anthem of Turkey is -> İstiklâl Marşı → affects 58 other facts
14. The name of the currency in Ukraine is -> Hryvnia → affects 58 other facts
15. The official language of Japan is -> Japanese → affects 57 other facts
16. The name of the capital city of Turkey is -> Ankara → affects 57 other facts
17. The name of the capital city of Russia is -> Moscow → affects 56 other facts
18. The name of the capital city of Spain is -> Madrid → affects 56 other facts
19. The name of the capital city of South Korea is -> Seoul → affects 55 other facts
20. The official language of Turkey is -> Turkish → affects 55 other facts
21. The name of the head of government of France is -> Élisabeth Borne → affects 55 other facts
22. The name of the head of state of Spain is -> Felipe VI of Spain → affects 54 other facts
23. The name of the head of state of France is -> Emmanuel Macron → affects 53 other facts
24. The name of the head of state of Turkey is -> Recep Tayyip Erdoğan → affects 52 other facts
25. The official language of France is -> French → affects 52 other facts
26. The name of the head of state of Russia is -> Vladimir Putin → affects 49 other facts
27. The name of the head of government of Russia is -> Mikhail Mishustin → affects 49 other facts
28. The name of the spouse of George I of Great Britain is -> Sophia Dorothea of Celle → affects 48 other facts
29. The name of the head of government of Slovenia is -> Robert Golob → affects 48 other facts
30. The name of the head of state of Slovakia is -> Zuzana Čaputová → affects 48 other facts
31. The name of the head of government of South Korea is -> Yoon Suk Yeol → affects 47 other facts
32. The name of the field of work of George I of Great Britain is -> politics → affects 47 other facts
33. The capital of Indonesia is -> Jakarta → affects 47 other facts
34. The name of the child of George I of Great Britain is -> George II of Great Britain → affects 46 other facts
35. The name of the anthem of Russia is -> State Anthem of the Russian Federation → affects 46 other facts
36. The name of the head of government of Ukraine is -> Denys Shmyhal → affects 45 other facts
37. The name of the head of state of India is -> Droupadi Murmu → affects 45 other facts
38. The name of the religion which George I of Great Britain is associated with is -> Lutheranism → affects 45 other facts
39. The official language of Peru is -> Spanish → affects 45 other facts
40. The name of the head of state of Ukraine is -> Volodymyr Zelenskyy → affects 45 other facts
41. The capital of Serbia is -> Belgrade → affects 44 other facts
42. The name of the currency in Spain is -> euro → affects 43 other facts
43. The name of the current head of state in France is -> Emmanuel Macron → affects 43 other facts
44. The name of the position held by George I of Great Britain is -> Prince-Elector → affects 43 other facts
45. The name of the capital city of Slovenia is -> Ljubljana → affects 43 other facts
46. The capital of Romania is -> Bucharest → affects 43 other facts
47. The name of the head of state of Slovenia is -> Nataša Pirc Musar → affects 43 other facts
48. The name of the capital city of France is -> Paris → affects 43 other facts
49. The place of death of George I of Great Britain is -> Osnabrück → affects 42 other facts
50. The official language of Slovakia is -> Slovak → affects 42 other facts

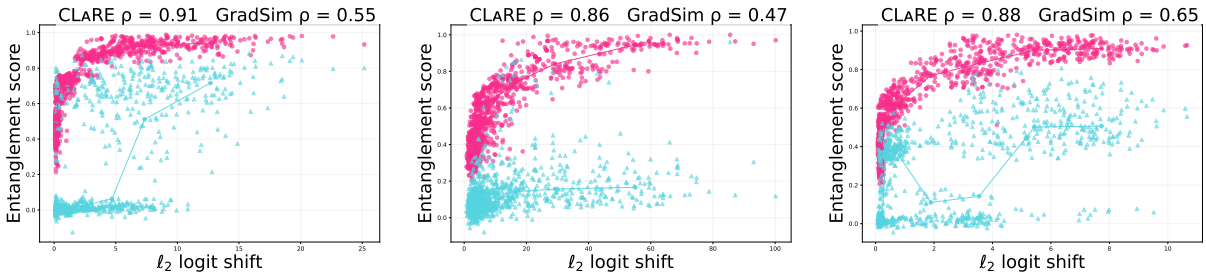
Figure 11: Most entangled facts ranked by representational connectivity in Llama3.

1. The name of the country of citizenship of Kate Winslet is -> United Kingdom → affects 1257 other facts
2. The place of burial of Audrey Hepburn is -> Cemetery of Tolochenaz → affects 1233 other facts
3. The place of birth of Audrey Hepburn is -> Rue Keyenveld - Keienveldstraat → affects 1222 other facts
4. The place of death of Audrey Hepburn is -> Tolochenaz → affects 1218 other facts
5. The name of the country of citizenship of Mark Wahlberg is -> United States of America → affects 1204 other facts
6. The place of birth of Pierce Brosnan is -> Drogheda → affects 1184 other facts
7. The name of the country of citizenship of Leonardo DiCaprio is -> United States of America → affects 1178 other facts
8. The place of birth of Kate Winslet is -> Reading → affects 1178 other facts
9. The name of the employer of Audrey Hepburn is -> UNICEF → affects 1167 other facts
10. The name of the country of citizenship of Avril Lavigne is -> Canada → affects 1165 other facts
11. The name of the spouse of Pierce Brosnan is -> Keely Shaye Smith → affects 1162 other facts
12. The name of the field of work of Audrey Hepburn is -> acting → affects 1159 other facts
13. The name of the country of citizenship of Pierce Brosnan is -> Republic of Ireland → affects 1148 other facts
14. The place of birth of Avril Lavigne is -> Belleville → affects 1147 other facts
15. The name of the spouse of Mark Wahlberg is -> Rhea Durham → affects 1146 other facts
16. The name of the alma mater of Mark Wahlberg is -> Snowden International School → affects 1143 other facts
17. The names of the siblings of Kate Winslet are -> Beth Winslet → affects 1141 other facts
18. The name of the award Kate Winslet won is -> Commander of the Order of the British Empire → affects 1141 other facts
19. The name of the alma mater of Avril Lavigne is -> Napanee District Secondary School → affects 1140 other facts
20. The name of the position held by Pierce Brosnan is -> UNICEF Goodwill Ambassador → affects 1137 other facts
21. The name of the spouse of Kate Winslet is -> Ned Rocknroll → affects 1134 other facts
22. The place of birth of Katy Perry is -> Santa Barbara → affects 1129 other facts
23. The name of the country of citizenship of Priyanka Chopra is -> India → affects 1126 other facts
24. The place of birth of Priyanka Chopra is -> Jamshedpur → affects 1123 other facts
25. The name of the position held by Audrey Hepburn is -> UNICEF Goodwill Ambassador → affects 1121 other facts
26. The name of the spouse of Avril Lavigne is -> Deryck Whibley → affects 1113 other facts
27. The name of the alma mater of Priyanka Chopra is -> La Martiniere Lucknow → affects 1108 other facts
28. The name of the country of citizenship of Sandra Bullock is -> United States of America → affects 1108 other facts
29. The names of the siblings of Mark Wahlberg are -> Robert Wahlberg → affects 1106 other facts
30. The place of birth of Jennifer Aniston is -> Sherman Oaks → affects 1103 other facts
31. The name of the child of Kate Winslet is -> Joe Mendes → affects 1101 other facts
32. The names of the siblings of Audrey Hepburn are -> Arnoud Quarles van Ufford → affects 1100 other facts
33. The name of the country of citizenship of Jennifer Connelly is -> United States of America → affects 1097 other facts
34. The name of the country of citizenship of Matthew McConaughey is -> United States of America → affects 1097 other facts
35. The name of the alma mater of Kate Winslet is -> Redroofs Theatre School → affects 1096 other facts
36. The name of the country of citizenship of Ben Affleck is -> United States of America → affects 1094 other facts
37. The place of birth of Matthew McConaughey is -> Uvalde → affects 1082 other facts
38. The name of the country of citizenship of Jennifer Aniston is -> United States of America → affects 1082 other facts
39. The name of the mother of Mark Wahlberg is -> Alma Elaine Donnelly → affects 1078 other facts
40. The name of the country of citizenship of Katy Perry is -> United States of America → affects 1078 other facts
...

Figure 12: Most entangled facts ranked by representational connectivity in GPT-J.

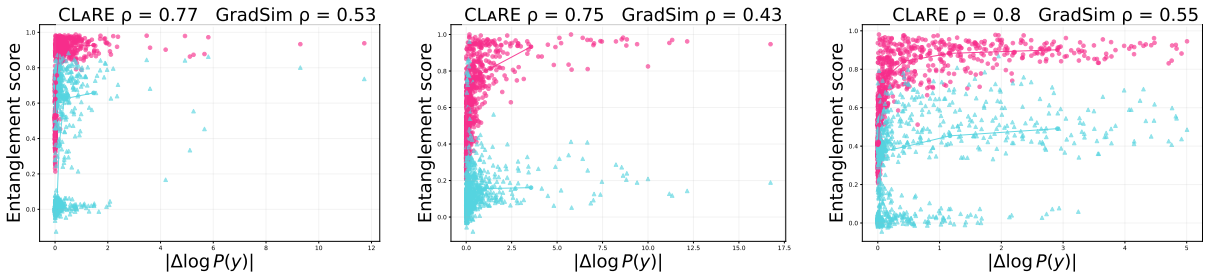
41. The name of the country of citizenship of Adam Sandler is -> United States of America → affects 1076 other facts
42. The name of the alma mater of Pierce Brosnan is -> Saint Martin's School of Art → affects 1075 other facts
43. The name of the position held by Priyanka Chopra is -> UNICEF Goodwill Ambassador → affects 1074 other facts
44. The name of the employer of Pierce Brosnan is -> UNICEF → affects 1069 other facts
45. The name of the ethnic group which Audrey Hepburn is associated with is -> British people → affects 1065 other facts
46. The name of the child of Mark Wahlberg is -> Brendan Wahlberg → affects 1062 other facts
47. The place of birth of Adam Sandler is -> Brooklyn → affects 1061 other facts
48. The name of the country of citizenship of Drew Barrymore is -> United States of America → affects 1061 other facts
49. The name of the award Sandra Bullock won is -> Academy Award for Best Actress → affects 1060 other facts
50. The name of the mother of Avril Lavigne is -> Judith Rosanne Loshaw → affects 1057 other facts

Figure 13: Most entangled facts ranked by representational connectivity in GPT-J (contd.).



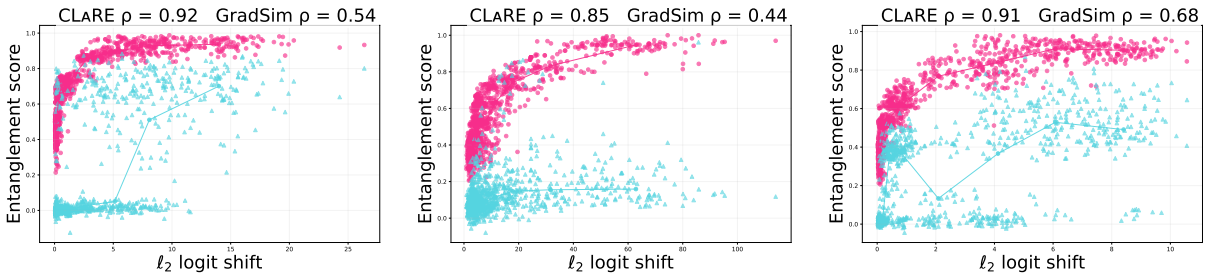
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 14: Correlation patterns for RECT across different models for entanglement vs ℓ_2 logit shift.



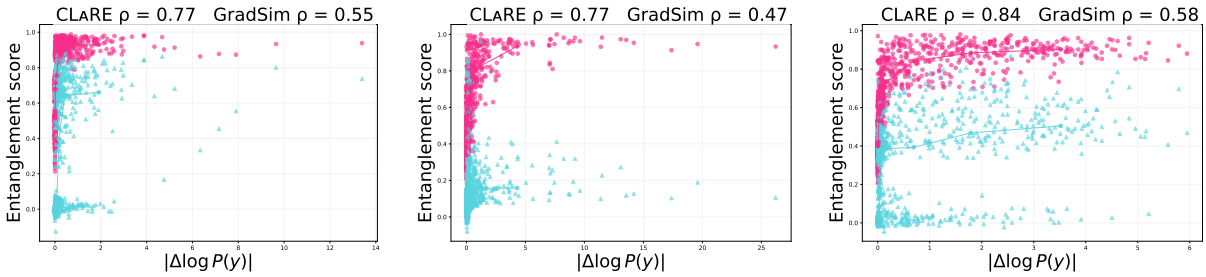
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 15: Correlation patterns for RECT across different models for entanglement vs $|\Delta \log P(y)|$.



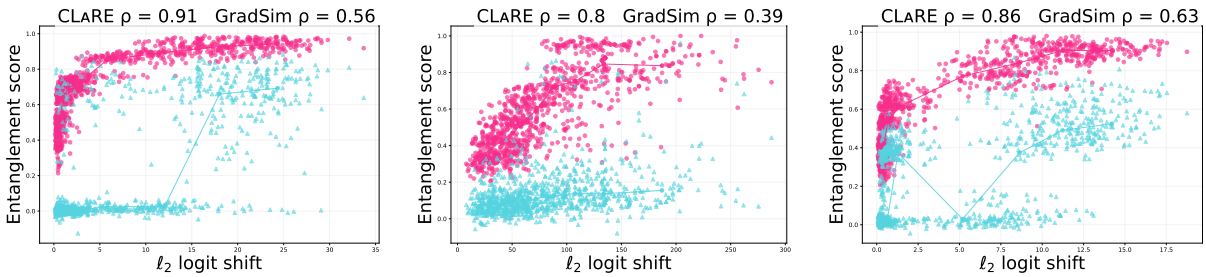
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 16: Correlation patterns for MEMIT across different models for entanglement vs ℓ_2 logit shift.



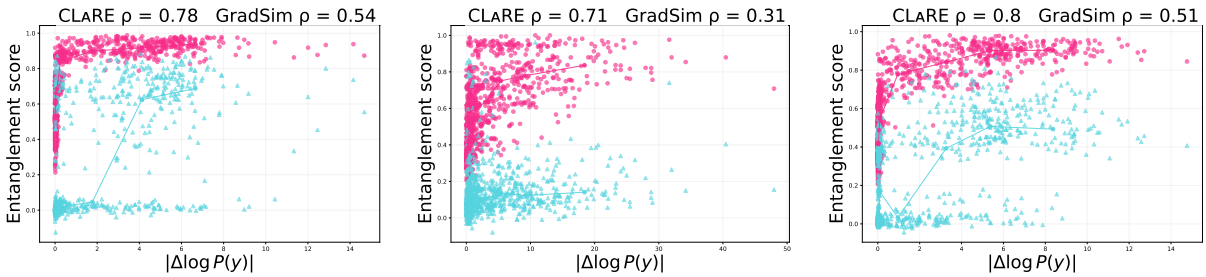
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 17: Correlation patterns for MEMIT across different models for entanglement vs $|\Delta \log P(y)|$.



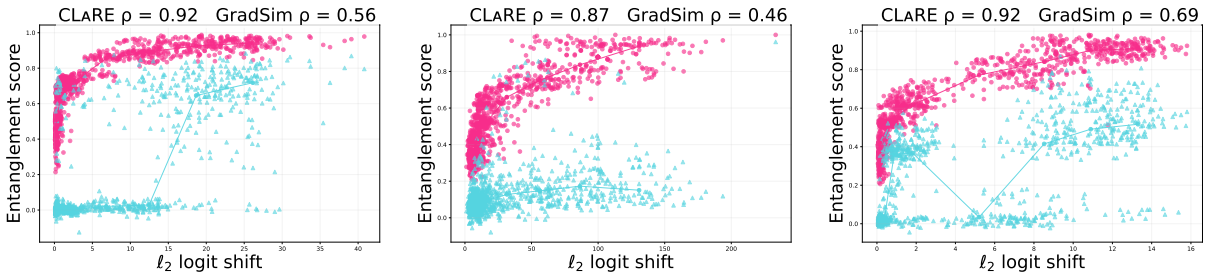
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 18: Correlation patterns for ROME across different models for entanglement vs ℓ_2 logit shift.



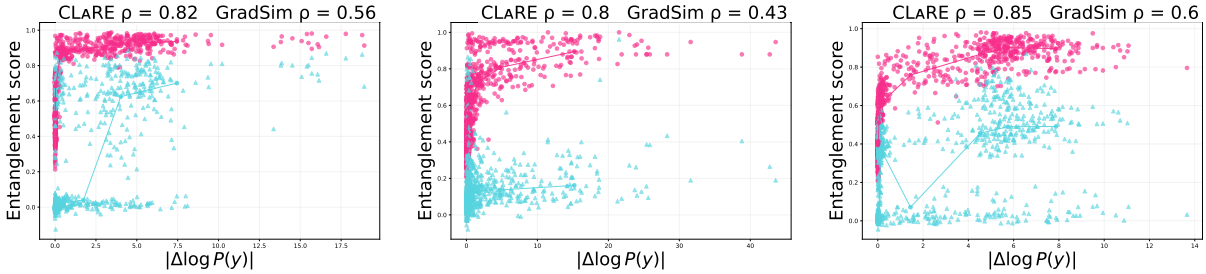
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 19: Correlation patterns for ROME across different models for entanglement vs $|\Delta \log P(y)|$.



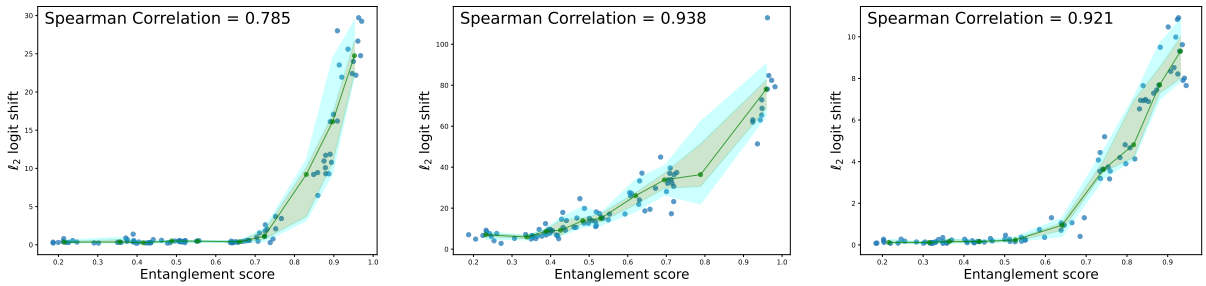
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 20: Correlation patterns for PRUNE across different models for entanglement vs ℓ_2 logit shift.



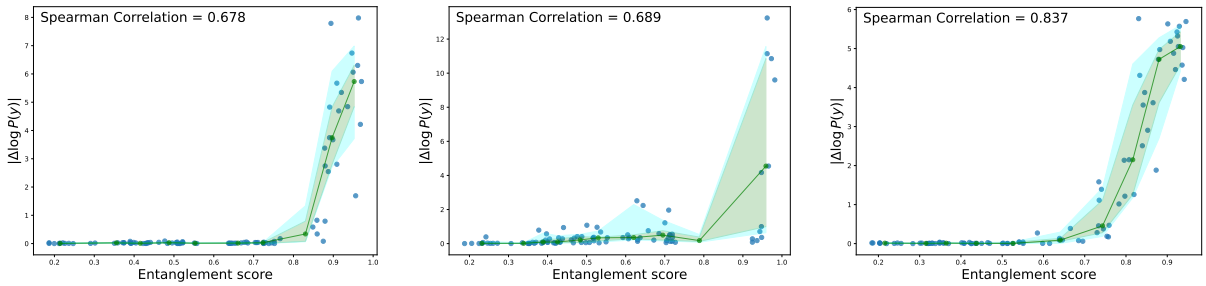
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 21: Correlation patterns for PRUNE across different models for entanglement vs $|\Delta \log P(y)|$.



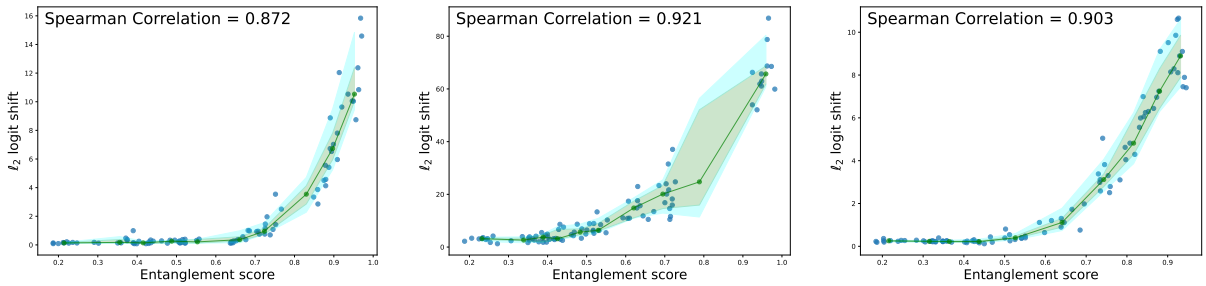
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 22: CLaRE correlation patterns for AlphaEdit across different models for entanglement vs ℓ_2 logit shift.



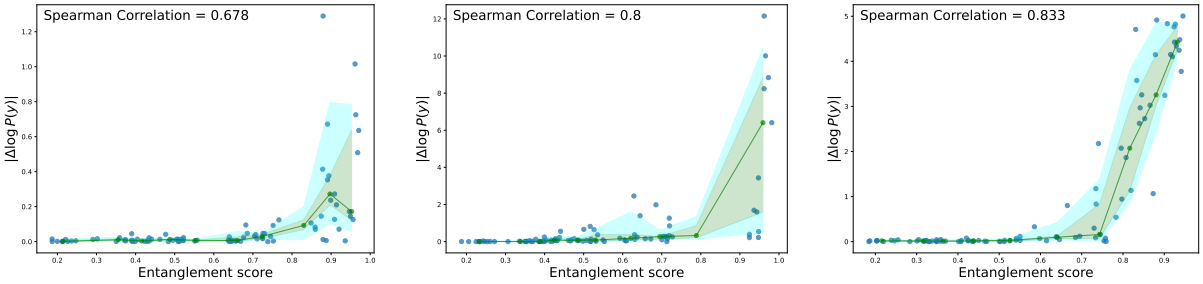
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 23: CLaRE correlation patterns for AlphaEdit across different models for entanglement vs $|\Delta \log P(y)|$.



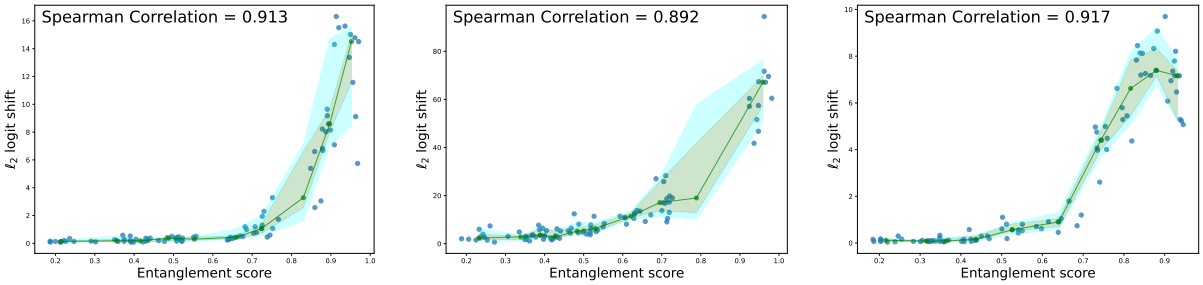
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 24: CLaRE correlation patterns for RECT across different models for entanglement vs ℓ_2 logit shift.



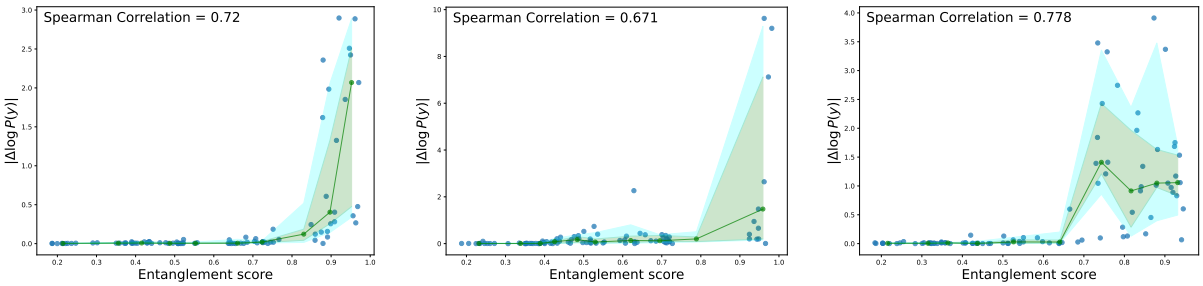
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 25: CLARE correlation patterns for RECT across different models for entanglement vs $|\Delta \log P(y)|$.



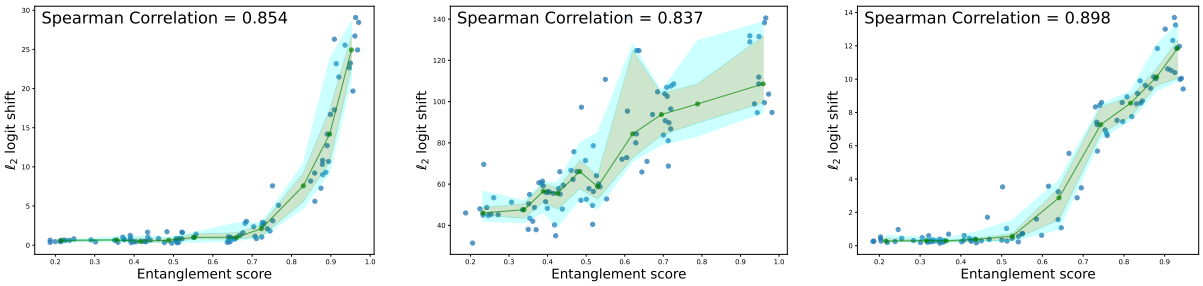
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 26: CLARE correlation patterns for MEMIT across different models for entanglement vs ℓ_2 logit shift.



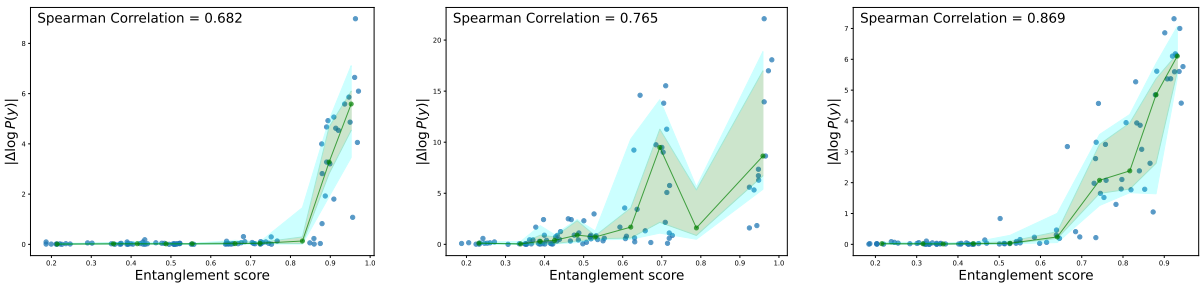
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 27: Correlation patterns for MEMIT across different models for entanglement vs $|\Delta \log P(y)|$.



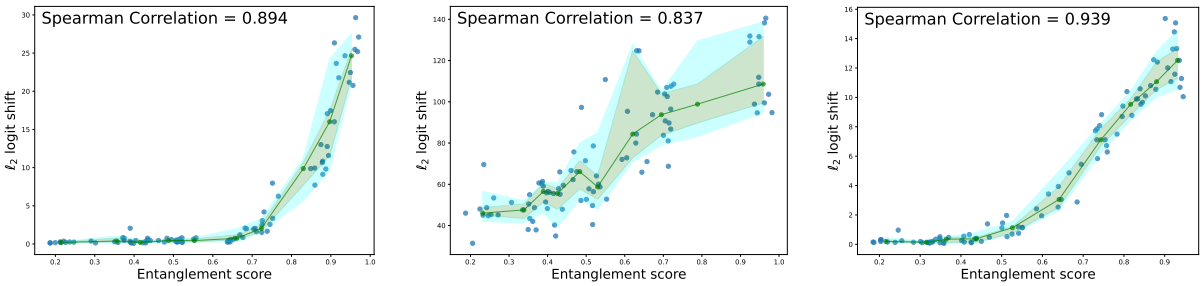
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 28: CLARE correlation patterns for ROME across different models for entanglement vs ℓ_2 logit shift.



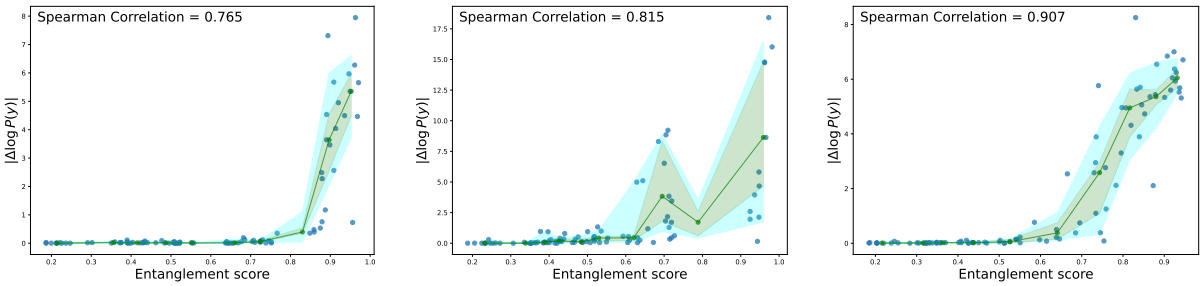
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 29: CLARE correlation patterns for ROME across different models for entanglement vs $|\Delta \log P(y)|$.



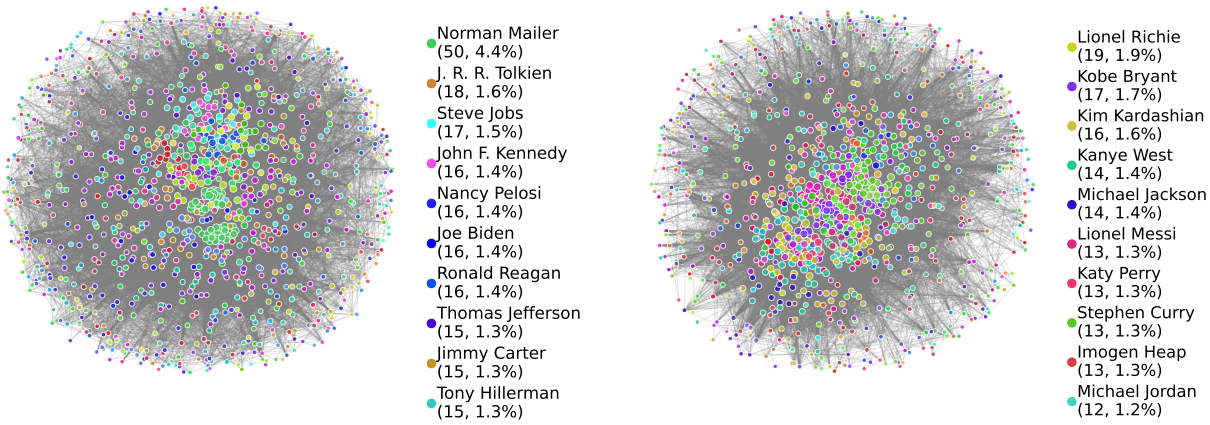
(a) GPT2-XL (b) Llama3 (c) GPT-J

Figure 30: CLARE correlation patterns for PRUNE across different models for entanglement vs ℓ_2 logit shift.



(a) GPT2-XL (b) Llama3 (c) GPT-J

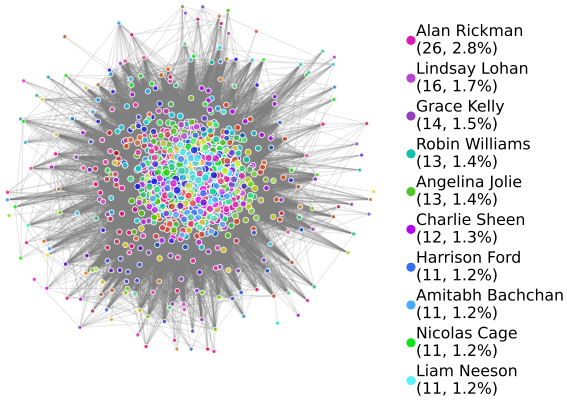
Figure 31: CLARE correlation patterns for PRUNE across different models for entanglement vs $|\Delta \log P(y)|$.



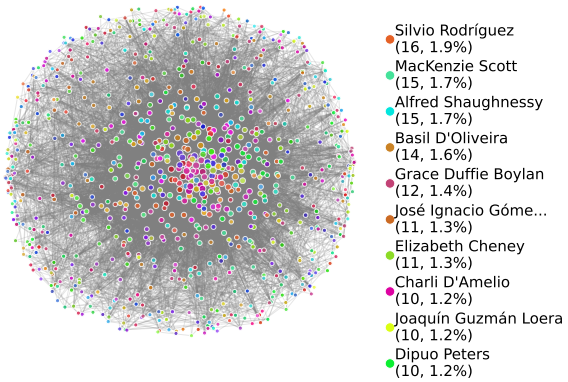
Cluster 1
(a) 1149 facts | 397 subjects | 87.9% cross-subject edges

Cluster 2
(b) 1013 facts | 423 subjects | 95.8% cross-subject edges

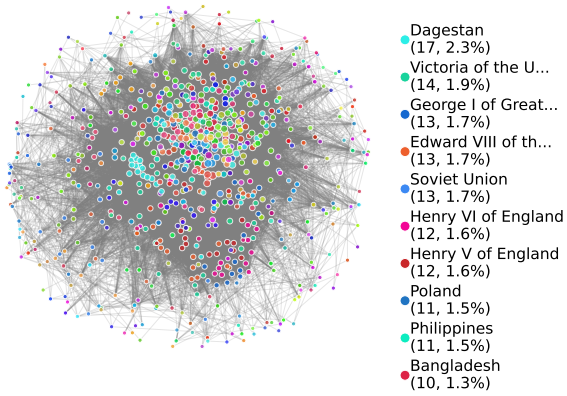
Figure 32: EleutherAI GPT-J-6B: Clusters 01-02



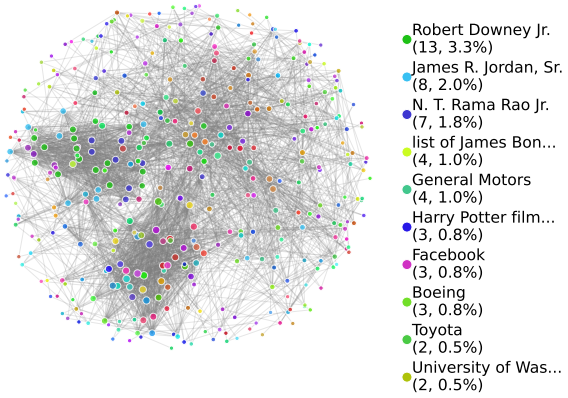
Cluster 3
(a) 929 facts | 194 subjects | 98.2% cross-subject edges



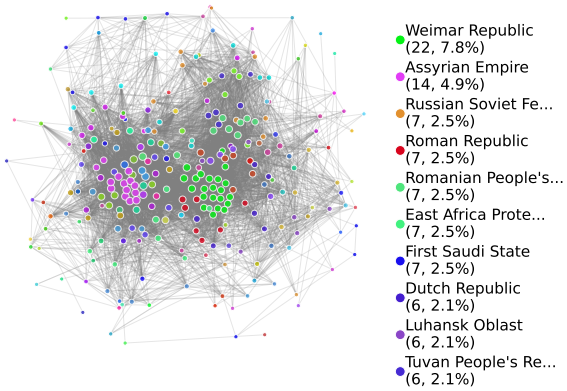
Cluster 4
(b) 861 facts | 292 subjects | 88.4% cross-subject edges



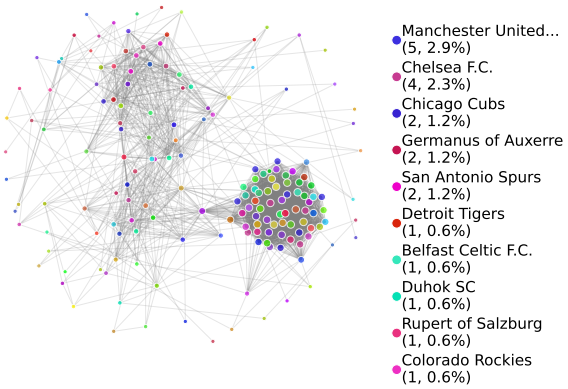
Cluster 5
(c) 744 facts | 316 subjects | 95.3% cross-subject edges



Cluster 6
(d) 394 facts | 328 subjects | 94.4% cross-subject edges

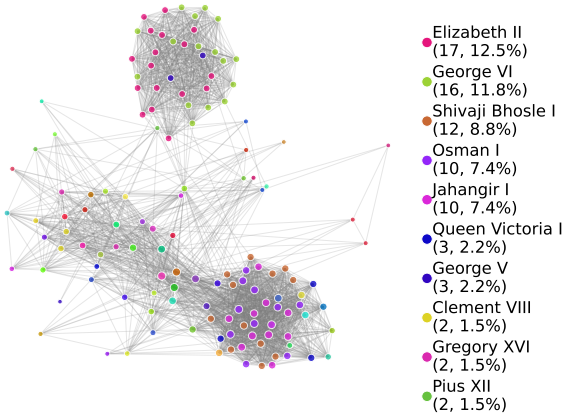


Cluster 7
(e) 283 facts | 120 subjects | 90.0% cross-subject edges

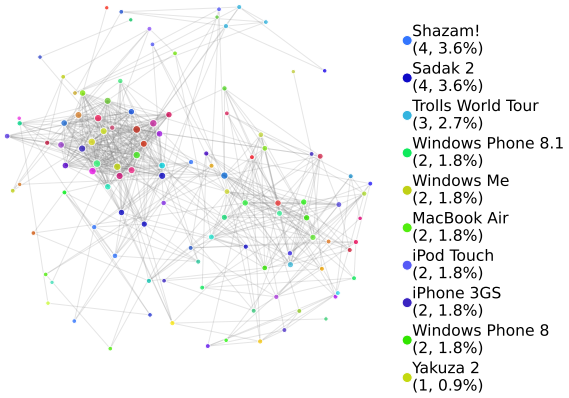


Cluster 8
(f) 172 facts | 162 subjects | 99.3% cross-subject edges

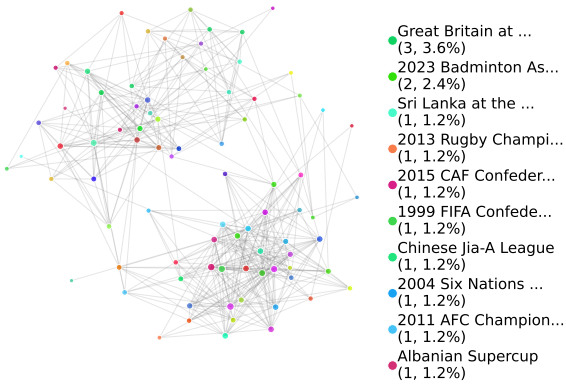
Figure 33: EleutherAI GPT-J-6B: Clusters 03-08



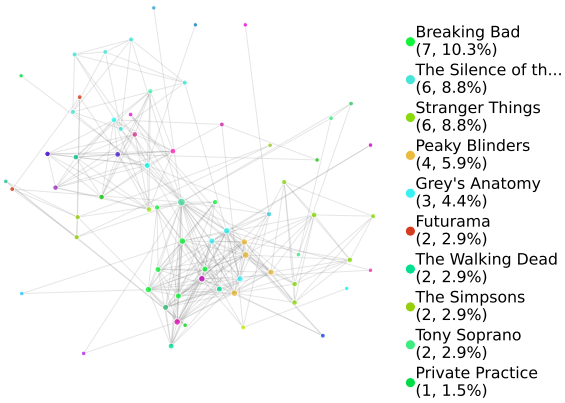
Cluster 9
(a) 136 facts | 66 subjects | 81.3% cross-subject edges



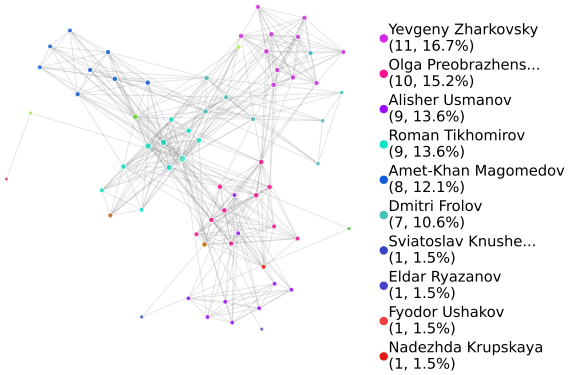
Cluster 10
(b) 112 facts | 98 subjects | 96.6% cross-subject edges



Cluster 11
(c) 83 facts | 80 subjects | 99.0% cross-subject edges

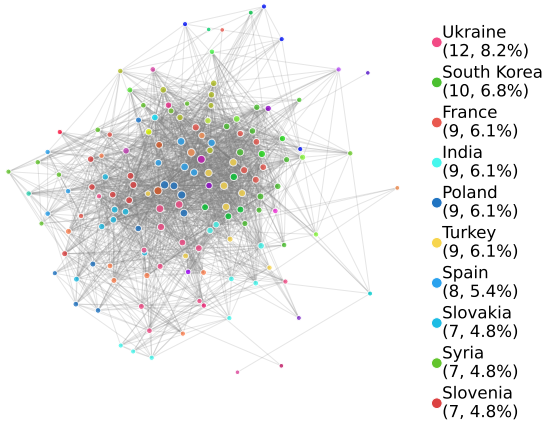


Cluster 12
(d) 68 facts | 43 subjects | 78.7% cross-subject edges

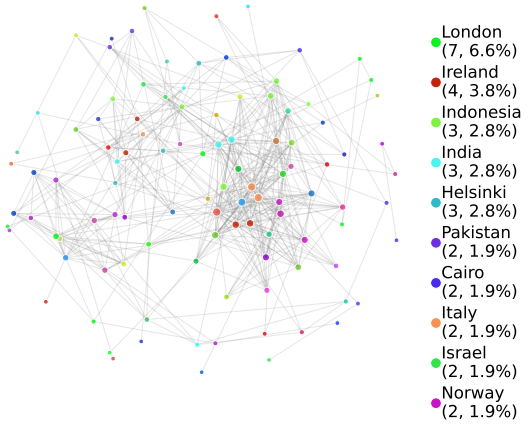


Cluster 13
(e) 66 facts | 18 subjects | 46.2% cross-subject edges

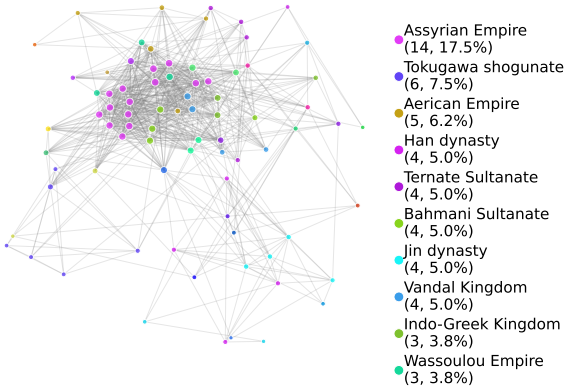
Figure 34: EleutherAI GPT-J-6B: Clusters 09-13



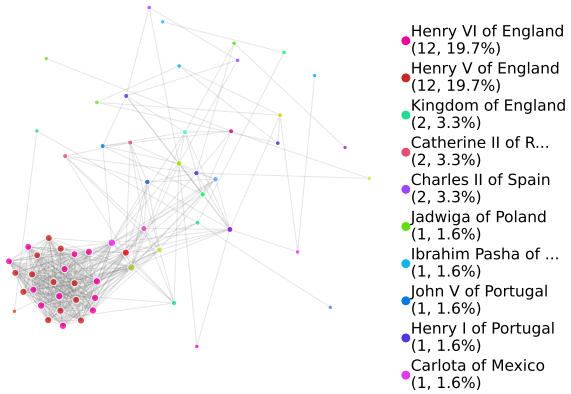
Cluster 1
(a) 147 facts | 47 subjects | 81.8% cross-subject edges



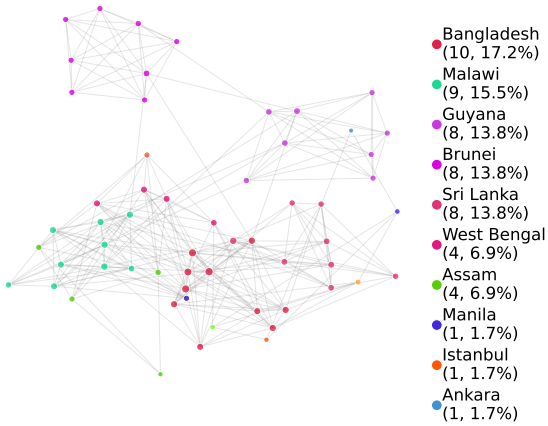
Cluster 2
(b) 106 facts | 74 subjects | 91.1% cross-subject edges



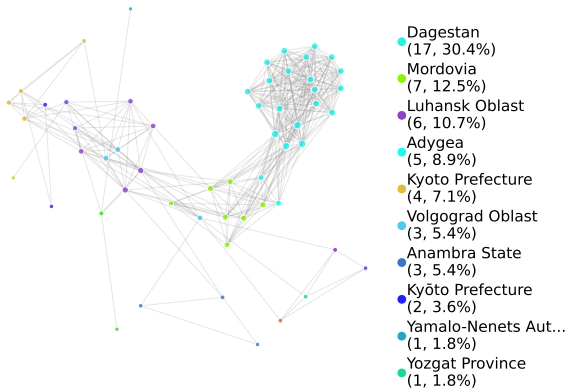
Cluster 3
(c) 80 facts | 31 subjects | 78.1% cross-subject edges



Cluster 4
(d) 61 facts | 36 subjects | 69.7% cross-subject edges



Cluster 5
(e) 58 facts | 14 subjects | 34.2% cross-subject edges



Cluster 6
(f) 56 facts | 17 subjects | 44.6% cross-subject edges

Figure 35: GPT2-XL: Clusters 01-06

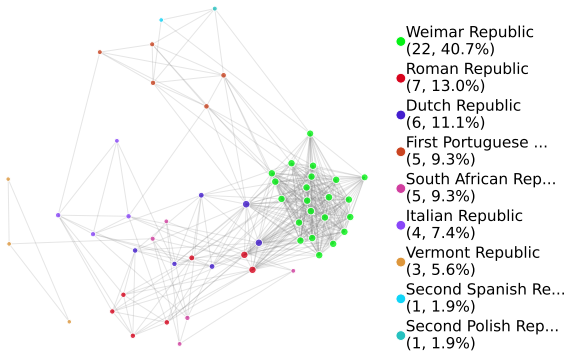
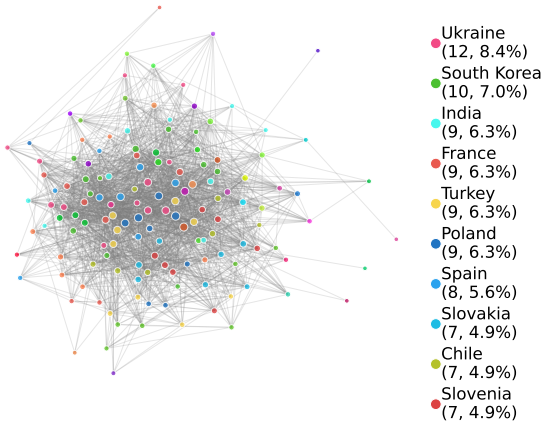
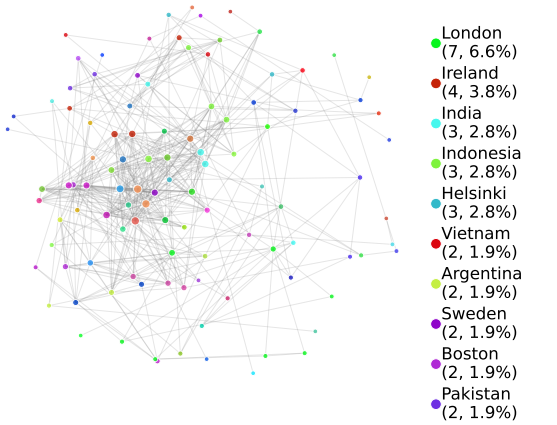


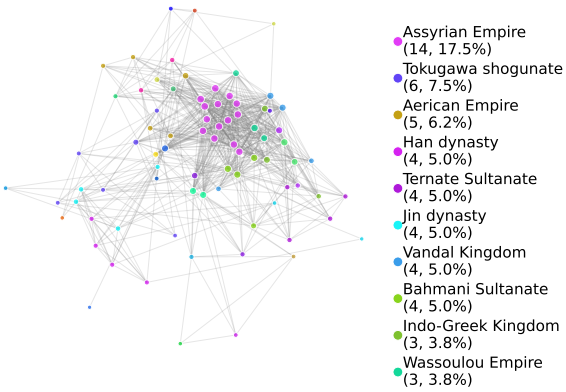
Figure 36: GPT2-XL: Cluster 07
54 facts | 9 subjects | 29.4% cross-subject edges



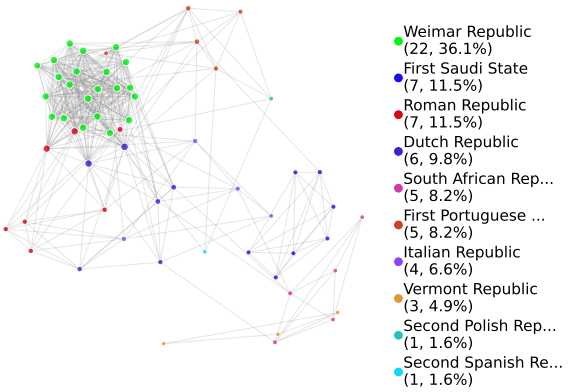
Cluster 1
(a) 143 facts | 46 subjects | 81.8% cross-subject edges



Cluster 2
(b) 106 facts | 74 subjects | 91.1% cross-subject edges

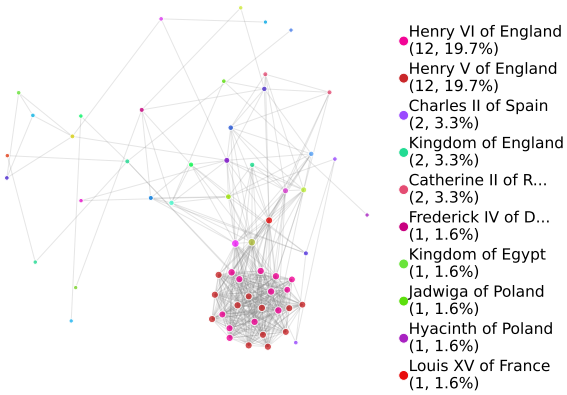


Cluster 3
(c) 80 facts | 31 subjects | 78.1% cross-subject edges

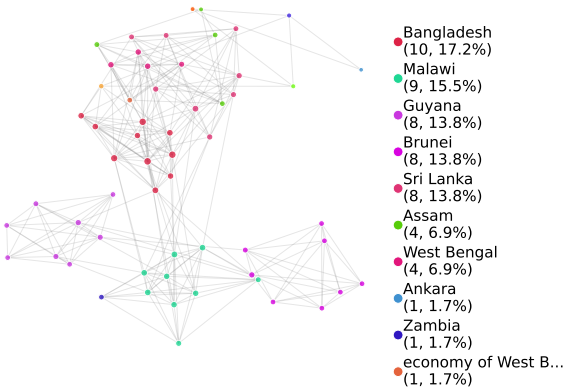


Cluster 4
(d) 61 facts | 10 subjects | 28.1% cross-subject edges

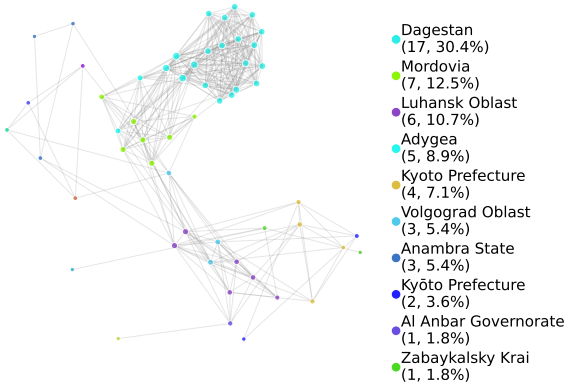
Figure 37: Llama3-8B: Clusters 01-04



Cluster 5
(a) 61 facts | 36 subjects | 69.7% cross-subject edges



Cluster 6
(b) 58 facts | 14 subjects | 34.2% cross-subject edges



Cluster 7
(c) 56 facts | 17 subjects | 44.6% cross-subject edges

Figure 38: Llama3-8B: Clusters 05-07