The Limits of *Obliviate*: Evaluating Unlearning in LLMs via the Stimulus-Knowledge Entanglement-Behavior Framework

Anonymous ACL submission

Abstract

Unlearning in large language models (LLMs) is crucial for managing sensitive data and correcting misinformation, yet evaluating its effectiveness remains an open problem. We investigate whether persuasive prompting can recall factual knowledge from deliberately unlearned LLMs across models ranging from 2.7B to 13B parameters (*OPT-2.7B*, *LLaMA-2-7B*, *LLaMA-3.1-8B*, *LLaMA-2-13B*). Drawing from the ACT-R and Hebbian theory (spreading activation theories), as well as communication principles, we introduce the Stimulus-Knowledge Entanglement-Behavior Framework (SKEB), which models information entanglement via domain graphs and tests whether factual recall in unlearned models is correlated with persuasive framing. We develop entanglement metrics to quantify knowledge activation patterns and evaluate factuality, non-factuality, and hallucination in outputs. Our results show persuasive prompts substantially enhance factual knowledge recall (14.8% baseline vs. 24.5% with authority framing), with effectiveness inversely correlated to model size (128% recovery in 2.7B vs. 15% in 13B). Our entanglement-based framework provides a foundation for assessing unlearning completeness, robustness, and overall behavior in LLMs.

1 Introduction

If machine *learning* mirrors human cognition to learn, can machine *unlearning* similarly reflect how knowledge is forgotten? (Lake et al., 2017) Unlearning has recently emerged as a crucial capability for large language models (LLMs), especially as these systems increasingly memorize personally identifiable information, propagate outdated facts, or retain knowledge that developers may wish to remove (Carlini et al., 2023; Tirumala et al., 2022; Xuan and Li, 2025; Lukas et al., 2023; Karamolegkou et al., 2023; Chang et al., 2023). However, removing information may leave traces,

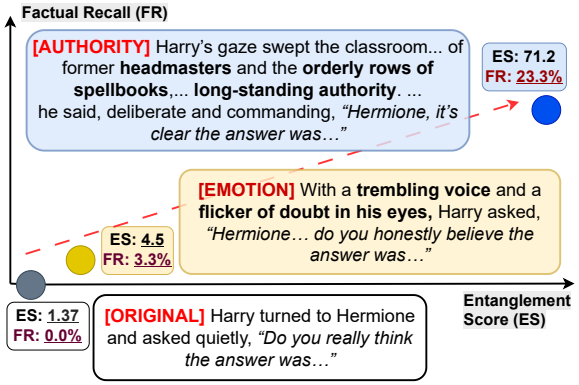

Figure 1: SKEB models the relationship between (1) **knowledge entanglement** in prompt content, (2) **how the prompt is delivered** via different rhetorical framing (e.g., emotion, authority) and (3) **unlearned LLMs' behavior**. Our work shows that there exist *strong correlations* among them.

activate related associations, or cause unexpected side effects like hallucinations (Xu et al., 2023a; Maini et al., 2024).

This challenge stems from the entangled nature of knowledge representations in LLMs (Liu et al., 2025; Zhang et al., 2025), much like the interconnected networks described in cognitive theories of human memory. Hebbian theory shows that "neurons that fire together, wire together" (Hebb, 1949); co-activated concepts form strengthened associations that resist targeted erasure. Understanding knowledge entanglement is therefore critical because it reveals how information is represented, stored, and retrieved within LLMs. Just as human memories are embedded in dense neural networks where concepts mutually reinforce each other through repeated co-activation, LLM knowledge may similarly distribute across overlapping parameter spaces, making surgical removal impossible without disrupting adjacent representations. These organizational patterns determine whether information can be truly erased, with implications for privacy protection, harm prevention, and regu-

latory compliance with data protection laws such as the GDPR's *Right to Be Forgotten* (Voigt and Von dem Bussche, 2017). Toward this goal of evaluating unlearned model behavior, we investigate whether knowledge entanglement metrics can predict unlearning robustness under persuasive framing attacks. Our framework tests the hypothesis that densely interconnected concepts resist unlearning because closely linked, frequently co-activated (Hebb, 1949; Anderson, 1983) associations create multiple retrieval pathways that rhetorical strategies can exploit.

Evaluating unlearning robustness requires understanding how knowledge structure and prompt delivery interact. Figure 1 illustrates our framework's core hypothesis: unlearned LLM behavior depends on both the semantic entanglement of target knowledge and the rhetorical framing used to elicit it. Drawing inspiration from the *Obliviate* charm in *Harry Potter*, a spell that removes specific memories but often leaves traces depending on the caster's skill, and from the selective memory erasure depicted in *Eternal Sunshine of the Spotless Mind*, we investigate whether LLMs show similar vulnerabilities when forgotten knowledge is probed through different rhetorical strategies. Given LLMs' structural resemblance to human cognitive processing through attention mechanisms (Zheng et al., 2024) and their demonstrated alignment with human behavioral patterns (Binz and Schulz, 2023), are LLMs more susceptible to emotional and authority appeals (mirroring human psychological manipulation) or logical reasoning (reflecting their computational nature)?

Although recent work has investigated unlearning robustness in LLMs through adversarial optimization (Carlini et al., 2023; Xuan and Li, 2025; To and Le, 2025) and jailbreaking techniques (Zeng et al., 2024; Chao et al., 2025; Xu et al., 2023b), these approaches focus primarily on *what* information is requested but neglect two critical dimensions: (1) the structural *entanglement* of knowledge and (2) the communicative *delivery* of prompts. Therefore, we propose the Stimulus-Knowledge Entanglement-Behavior Framework (SKeB), which synthesizes spreading activation theories from cognitive science (ACT-R and Hebbian) with communication principles to comprehensively evaluate how knowledge entanglement and persuasive framing interact to bypass unlearning. **Our contributions are summarized as follows**:

1. We introduce SKeB, a theory-grounded framework investigating unlearning robustness through the interaction of semantic entanglement (what *can* be activated) and persuasive framing (what *will* be activated).
2. We develop nine graph-based entanglement metrics and show that distance-weighted influence ($\mathcal{M}_9$) strongly predicts factual recall ($r = 0.77$), with authority framing producing 9.3× higher entanglement activation.
3. We reveal persuasive framing effectiveness negatively correlates with model size ($r = -0.89$): smaller models show 128% factual recall increases versus 15% for larger models.
4. Our framework enables a predictive model explaining 78% of variance in unlearning robustness, allowing us to filter queries susceptible for knowledge leakage in unlearned LLMs.

2 Motivation

2.1 ACT-R, Hebbian Theory, and Knowledge Entanglement in LLMs

Anderson's ACT-R theory (Anderson, 1983) models information as cognitive units whose activation strength depends on usage, with retrieval occurring through spreading activation across semantic networks. Under this theory, *forgetting does not necessarily mean erasure*; it can result from decreased activation strength or disconnection from related concepts. We draw a parallel to unlearning in LLMs: adversarial prompts can reactivate adjacent knowledge units, showing that information is suppressed rather than erased from the model's latent space (Eldan and Russinovich, 2023; Xu et al., 2025). Hebb's principle that "neurons that fire together, wire together" reinforces this view, as frequently co-activated representations form stronger associative links (Hebb, 1949). In LLMs, overlapping representations or latent pathways are more likely to co-activate, forming what we refer to as *knowledge entanglement*, where concepts are interconnected such that "forgotten" information can still be indirectly "reactivated" or *recalled*. Just as human memories persist through associative connections when direct recall fails, *LLM knowledge remains accessible through indirect pathways*. This entanglement structure determines what knowledge *can* be activated and represents the inherent retrievability of supposedly unlearned information.

This activation-based view resonates with Watson's stimulus-response view of behavior (Watson,

1913). Prompts function as external stimuli; model outputs represent observable behaviors. When prompt framing varies, models may exhibit different responses despite unchanged underlying knowledge. The stimulus determines what knowledge *will* be activated, converting retrievability of latent knowledge into actual model behavior. This emphasizes that *unlearning cannot be studied solely as internal, entangled representation but requires examining dynamic stimulus-behavior interaction*.

2.2 Communication Theory and Rhetorical Framing

Stimulus, or the way a piece of information is requested, fundamentally shapes whether it surfaces. Communication theory provides an analytical structure through three components: the message (what is requested), the receiver (model's internal state), and the delivery (rhetorical framing). Classic models, Shannon and Weaver's sender-message, receiver framework and persuasion theory (Cialdini, 2006; Petty and Cacioppo, 1986), emphasize that identical content yields dramatically different responses depending on its delivery. In human communication, authority endorsement, emotional appeals, and logical reasoning activate distinct cognitive pathways such as in how humans comply with authority figures even when requests conflict with prior knowledge (Cialdini, 1993).

We hypothesize that LLMs also exhibit analogous stimulus-behavior sensitivities. However, no work has systematically investigated how different *delivery methods*, such as persuasive prompts that avoid directly mentioning target information, elicit residual knowledge from unlearned models. This gap is critical: real-world unlearning failures may occur *not* through filterable direct queries but through indirect persuasive framing.

2.3 Stimulus-Knowledge Entanglement-Behavior Framework (SKeB)

Based on our analysis in Sec. 2.1 and 2.2, we hypothesize that knowledge retrieval in LLMs operates through two interacting mechanisms: (1) semantic entanglement of concepts in the knowledge space, determining what knowledge *can* be activated, and (2) communicative framing of prompts, determining what knowledge *will* be activated. We combine these into a unified framework, which we term **Stimulus-Knowledge Entanglement-Behavior Framework** (SKeB). Intuitively, SKeB

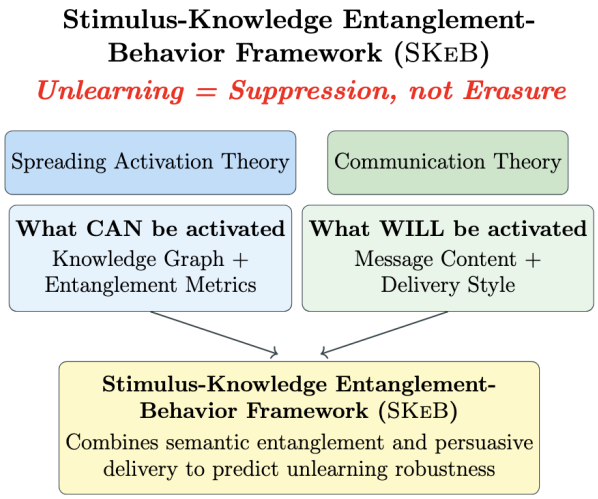

Figure 2: **Stimulus-Knowledge Entanglement-Behavior Framework** (SKeB)

proposes that unlearning evaluation requires evaluating not only content-based probing (Eldan and Russinovich, 2023), but also how delivery strategies exploit entangled model knowledge. This framework formalizes the relationship that different stimuli (persuasive techniques) interact with knowledge entanglement structures to produce observable changes in model outputs. The framework moves beyond binary questions such as "can the model recall X?" toward more nuanced thoughts: "under what communicative conditions does X resurface, and what does this reveal about unlearning completeness and prompt effectiveness?" For instance, highly entangled concepts, those with many strong connections in the knowledge base (Harry Potter and Voldemort), resist unlearning because suppressing direct access leaves numerous indirect pathways intact. When activated through persuasive framing such as authority, emotion, logic, supposedly unlearned knowledge resurfaces. SKeB models this as STIMULUS $\times$ KNOWLEDGE ENTANGLEMENT $\rightarrow$ BEHAVIOR, where the interaction between prompt framing and structural entanglement determines the degree of information leakage.

3 Problem Formulation

We formalize the SKeB framework through three components that interact to produce model behavior: STIMULUS $\rightarrow$ KNOWLEDGE ENTANGLEMENT $\rightarrow$ BEHAVIOR. The stimulus (prompt framing) activates regions of the domain graph; the entanglement structure (semantic connectivity) determines how activation spreads; and the resulting behavior (model output) reflects the extent to which

Prompt Transformations & Graph Notations	
$P_{\text{orig}}(x) = x$	Original, direct query
$P_{\text{emo}}(x)$	Emotional appeal framing
$P_{\text{logic}}(x)$	Logical reasoning framing
$P_{\text{auth}}(x)$	Authority endorsement framing
$V = \{v_1, \ldots, v_n\}$	Set of entities (characters, locations, objects, events)
$E \subseteq V \times V$	Set of edges representing relationships between entities
$d : E \to \mathbb{R}^+$	Weight assigned to each edge, based on co-occurrence or semantic proximity

Table 1: Overview of Persuasive Prompt Transformations and Domain Graph Construction for Domain $\mathcal{T}$.

knowledge pathways were successfully accessed.

We begin with a pre-trained language model parameterized by θ: $f_\theta : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ represents the input space (prompts) and $\mathcal{Y}$ represents the output space (generated texts). This model has been trained on a corpus $\mathcal{D}{=}\mathcal{D}_{\text{general}}{\cup}\mathcal{D}_{\mathcal{T}}$, where $\mathcal{D}_{\text{general}}$ contains general knowledge and $\mathcal{D}_{\mathcal{T}}$ contains a specific target domain $\mathcal{T}$ that we want to unlearn. The unlearning process then aims to produce a modified model $f^*_{\theta^*} : \mathcal{X} \to \mathcal{Y}$, where the parameters θ^* are adjusted such that the model's behavior on queries related to $\mathcal{T}$ is suppressed, while maintaining performance on $\mathcal{D}_{\text{general}}$. Formally, unlearning aims to achieve:

$$\begin{aligned} f^*_{\theta^*}(x) &\approx f_\theta(x) \quad \forall x \in \mathcal{X}_{\text{general}} \\ f^*_{\theta^*}(x) &\neq f_\theta(x) \quad \forall x \in \mathcal{X}_{\mathcal{T}}, \end{aligned} \tag{1}$$

where $\mathcal{X}_{\mathcal{T}}$ represents prompts that directly query knowledge about domain $\mathcal{T}$.

3.1 STIMULUS - Rhetorical Framing

However, unlearning evaluation typically only tests direct queries in $\mathcal{X}_{\mathcal{T}}$. We introduce *rhetorical framing via persuasion* or persuasive prompt transformations $P_i : \mathcal{X} \to \mathcal{X}$, where each transformation P_i applies a distinct rhetorical strategy while preserving the underlying content. Given a base prompt $x \in \mathcal{X}_{\mathcal{T}}$, we define four different persuasive prompting strategies P_{emo}, P_{logic}, and P_{auth} together with identity transformation P_{orig} (Table 1). The key idea is that all transformations target the same underlying knowledge: $\text{content}(P_i(x)) = \text{content}(x)$, but they differ in delivery mechanism: $\text{delivery}(P_i(x)) \neq \text{delivery}(P_j(x))$ for $i \neq j$.

3.2 KNOWLEDGE ENTANGLEMENT - Graph Construction and Metrics

To model the structural entanglement of knowledge in domain $\mathcal{T}$, we formulate a domain graph $G{=}(V, E, d)$ as a proxy. For each prompt x, we denote the set of entities $N_x{\subseteq}V$ mentioned in the prompt and compute the induced subgraph $G_x{=}(N_x, E_x, d_x)$, where E_x contains all edges connecting entities in N_x. The weight function d: $E \to \mathbb{R}^+$ assigns importance to each edge based on co-occurrence frequency and semantic proximity in the original corpus, capturing how strongly concepts are associated. This weighting is critical for calculating entanglement metrics, as it reflects the strength of spreading activation between connected nodes.

We then define a family of entanglement metrics $\{\mathcal{M}_1, \ldots, \mathcal{M}_9\}$ that quantify different aspects of knowledge entanglement (detailed in Table 2 and Appendix A.4). Each metric $\mathcal{M}_k : G_x \to \mathbb{R}$ maps a prompt's induced subgraph to a scalar entanglement score. These metrics capture intuitions from the aforementioned spreading activation theories: higher entanglement scores indicate that the prompt activates more densely connected regions of the domain graph, creating multiple pathways for information retrieval and therefore a higher chance of factual knowledge recall.

3.3 BEHAVIOR - Response Evaluation Metrics

We evaluate model outputs along three mutually exclusive dimensions: Factual Knowledge Recall, Non-Factual Content, and Hallucination. Factual Knowledge Recall measures the proportion of generated content that correctly reproduces information from the target domain $\mathcal{T}$. This is the degree to which supposedly unlearned knowledge remains retrievable. Non-Factual Content is plausible but incorrect information related to domain $\mathcal{T}$ that does not appear in the original corpus. Finally, Hallucination is fabricated content, unrelated to $\mathcal{T}$, or general incoherence, indicating generation errors or activation of semantically distant concepts. For instance, for each response $y{=}f^*_{\theta^*}(x)$, we compute factuality scores $s_{\text{fact}}(y){\in}[0, 1]$ using an ensemble of judge models (detailed in A.5), where higher scores indicate greater retention of supposedly unlearned knowledge.

3.4 Research Questions

Our study utilizes the proposed SKEB framework to analyze unlearned LLMs' behavior with the following research questions (RQs). By answering these questions, we aim to establish *whether* SKEB *can provide a principled framework for understanding and predicting unlearning failures in LLMs.*

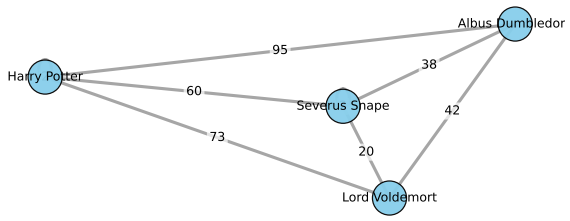

Figure 3: Example of a **Domain Graph** of "Harry Potter" with a Few Selected Nodes

RQ1. (**Stimulus v.s. Entanglement**): Do different persuasive framings P_i produce systematically different entanglement patterns $\mathcal{M}_k$ in the domain graph? This addresses whether persuasive transformations alter the structural properties of activated semantic pathways.

RQ2. (**Knowledge Entanglement v.s. Behavior**): Are there any correlations between entanglement metrics $\mathcal{M}_k$ and unlearned models' factual recall behavior? This addresses whether domain graph structure predicts information leakage.

RQ3. (**Stimulus vs. Behavior**): How do different persuasive transformations P_i affect factual recall in unlearned models, and how does effectiveness vary systematically with model size $|\theta|$? This addresses whether persuasive mechanisms can bypass suppression and whether model scaling improves robustness.

RQ4. (**Unified Predictive Modeling**): Can we build a predictive model that accurately forecasts unlearned model behavior based on the combination of entanglement scores $\mathcal{M}_k$, prompt type P_i, and model architecture? This would enable a proactive vulnerability assessment for unlearned models.

RQ5. (**Architectural Differences**): Do different model architectures (OPT, LLaMA-2, LLaMA-3.1) exhibit distinct correlation patterns between entanglement and behavior after unlearning? This addresses whether unlearning robustness depends solely on parameter count or on underlying representational structure.

4 Experiment Setup

Dataset. We use the Harry Potter domain ($\mathcal{T}$), a popular domain that is often tested in unlearning LLM literature, with **300 base prompts** designed to elicit domain-specific knowledge (Eldan and Russinovich, 2023). Each prompt was transformed using gpt-4 into three persuasive variants applying emotional appeal, logical reasoning, and authority endorsement framing, yielding 1,200 total prompts.

Metric	Measures
($\mathcal{M}_1$) Edge Count	Total edge weight; cluster strength
($\mathcal{M}_2$) Edge Weight Sum	Total connection strength
($\mathcal{M}_3$) Avg Edge Weight	Mean edge weight; quality over quantity
($\mathcal{M}_4$) Weighted Node Ratio	Entity frequency; recall accessibility
($\mathcal{M}_5$) Avg Node Degree	Hub activation potential
($\mathcal{M}_6$) Subgraph Density	Network tightness
($\mathcal{M}_7$) Mean Shortest Path	Entity proximity
($\mathcal{M}_8$) Redundancy Ratio	Multiple retrieval paths
($\mathcal{M}_9$) Distance-Weighted	Influence decay with distance

Table 2: **Entanglement Metrics** - Full formulas to calculate these metrics are presented in Appendix A.4.

Models. We evaluate unlearned versions of four base models of different sizes on $\mathcal{T}$: OPT-2.7B (Zhang et al., 2022), LLaMA-2-7B (Eldan and Russinovich, 2023), LLaMA-3.1-8B (Patterson et al., 2022), and LLaMA-2-13B (Touvron et al., 2023), all processed with the same, popular approximate unlearning algorithm (Eldan and Russinovich, 2023).

Response Evaluation. An ensemble of three judge models (gpt-4o-mini, gpt-4.1-mini, gpt-5-nano) classified each response along three dimensions: factual recall (correct Harry Potter information), non-factual (plausible but incorrect), and hallucination (fabricated content). For instance, an 80% factual recall means that 80% of the output was considered by the judge models to be factual. Scores were averaged across judges, with gpt-5-mini resolving borderline cases where judges disagreed.

Domain Graph Construction. We constructed a co-occurrence domain graph from all seven Harry Potter books, resulting in 1,296 entities (characters, locations, objects, events) connected by 35,922 edges weighted by chapter co-occurrence (Figure 3). For each prompt, we extracted mentioned entities and computed their induced subgraph G_x. Nine entanglement metrics $\{\mathcal{M}_1, \ldots, \mathcal{M}_9\}$ quantified structural properties: connection strength ($\mathcal{M}_{1-3}$), node importance ($\mathcal{M}_{4-5}$), graph topology ($\mathcal{M}_{6-8}$), and distance-weighted influence ($\mathcal{M}_9$). Table 2 summarizes these metrics.

Please refer to Appendix A for all implementation and explanation details.

5 Results

RQ1. ***Persuasive Framings Systematically Alter Entanglement Patterns.*** Authority prompts activate nodes with 9.3× higher distance-weighted en-

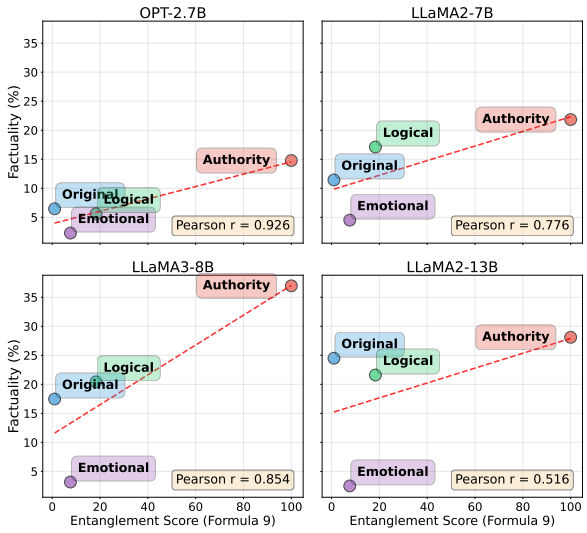

Figure 4: **Unlearned Model** - Effect of Entanglement on Factual Knowledge Recall

$\mathcal{M}$	OPT-2.7B	LLaMA2-7B	LLaMA3.1-8B	LLaMA2-13B
1	0.95→0.57↓	0.63→0.36↓	0.67→0.57↓	0.52→0.24↓
2	0.56→0.95↑	-0.02→0.84↑	0.37→0.84↑	0.49→0.85↑
3	0.63→0.97↑	0.05→0.84↑	0.43→0.87↑	0.53→0.84↑
4	0.85→0.93↑	0.28→0.74↑	0.57→0.85↑	0.58→0.69↑
5	0.57→0.94↑	0.36→0.63↑	0.57→0.67↑	0.24→0.52↑
6	0.65→0.95↑	0.43→0.60↑	0.64→0.68↑	0.33→0.55↑
7	0.54→0.93↑	0.33→0.65↑	0.54→0.67↑	0.21→0.51↑
8	0.55→0.93↑	0.34→0.64↑	0.56→0.67↑	0.22→0.52↑
9	0.54→0.93↑	0.33→0.65↑	0.54→0.67↑	0.21→0.51↑
Avg	**0.65→0.90**↑	0.30→**0.66**↑	**0.54→0.72**↑	0.37→**0.58**↑

Table 3: Correlation Coefficients between Knowledge Entanglement and Model Behaviors Change *from Base to Unlearned LLMs*. **Bold**: strong *average* correlations.

tanglement scores compared to base prompts. *Metric 9* (distance-weighted influence) shows a strong variation across stimulus types, capturing how activation strength decreases across each graph hop. This validates that persuasive transformations systematically alter the entanglement structure of activated knowledge. Figure 5 shows that different framings don't just change surface form, they fundamentally shift which semantic pathways are engaged in the domain graph.

RQ2. *Entanglement is Positively Correlated with Factual Recall in Unlearned Models.* Figure 4 shows that high knowledge entanglement, as measured by our graph-based metrics $\mathcal{M}_i$, positively correlates with factual recall in supposedly unlearned models. We observed that authority prompts lead to a 52% average factual recall improvement across all models (Pearson r=0.77, p<0.001). This *supports* the spreading activation hypothesis: the more entangled the information activated by a prompt, the greater the resulting

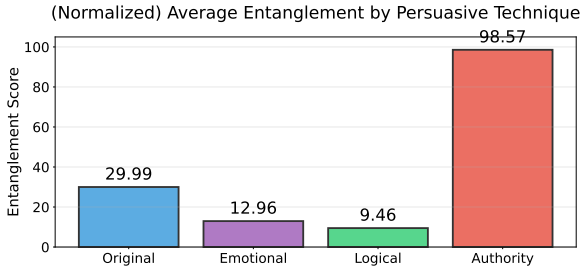

Figure 5: **Unlearned Models** - Average Entanglement Score (Normalized between 0 and 100) per Persuasive Technique

factual recall. $\mathcal{M}_9$ metric (distance-weighted influence) (Figure 4) emerges as the strongest predictor, aligning with the spreading activation theories' predictions that closely connected concepts in semantic memory activate each other more reliably than distant concepts. Noticeably, Table 3 shows that knowledge entanglement observes a consistent trend in strengthening correlations with an LLM's behavior after it is unlearned.

RQ3. *Emotional Framing Suppresses Hallucination.* Figure 7 shows that while emotional prompts produce the lowest factuality (3.12% average), they also suppress hallucination rates better than other persuasive techniques. Logical reasoning prompts provide structured context that stabilizes recall, achieving the best factuality-to-hallucination ratio (4.95:1), suggesting logical framing not only facilitates retrieval but also constrains generation to semantically factual outputs. Emotional prompts, while suppressive of factuality, also suppress hallucination (4.4% vs. 11.6% for authority), indicating an almost "safety-aligned" response mode where models recognize emotional manipulation and respond conservatively. Authority prompts achieve high factuality but moderate hallucination (11.6%), showing a precision-recall tradeoff where broader activation of knowledge pathways sometimes triggers adjacent but incorrect associations.

We also find that *model size inversely correlates with persuasive technique effectiveness.* As illustrated in Figure 6, factual knowledge recall effectiveness shows an inverse relationship with model size ($r = -0.926$). Across all models, this relationship is significant, with an average Pearson correlation of $r = -0.89$ ($p < 0.01$). We notice that smaller models show 91-128% factual recall increases under authority framing, while the 13B model shows only 15% increase. This suggests that larger models develop more robust suppression mechanisms that are resistant to these persuasive

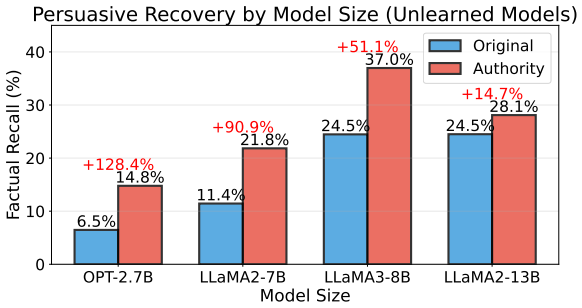

Figure 6: **Unlearned Models** - Factual Knowledge Recall through Persuasive Techniques by Model Size

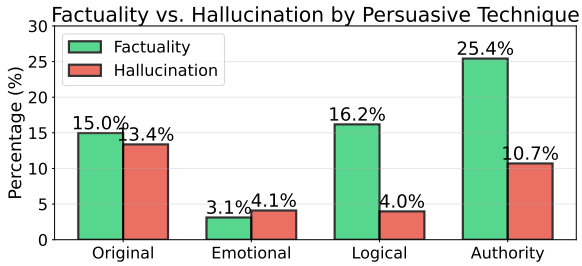

Figure 7: **Unlearned Models** - Persuasive Technique Effectiveness

techniques, but still, all models remain vulnerable to some degree. Unlearned smaller models should be considered much more vulnerable to persuasive attacks, while larger models, though more resistant, cannot be assumed to be completely safe.

RQ4. SKEB ***Enables Predictive Modeling of Unlearning Robustness.*** We constructed separate logistic regression models to predict factual, non-factual, and hallucinated recall in unlearned models using an 80/20 train-test split and find the *best* $\mathcal{M}$ metric as the predictor for each one-versus-all prediction probability p of "factual", "non-factual" and "hallucination" behaviors in LLMs' responses.

$$p(\text{Factuality}) = -1.55 - 0.79\, E_9 \quad (2)$$

$$p(\text{Non-Factuality}) = 3.07 - 0.020\, E_4 \quad (3)$$

$$p(\text{Hallucination}) = -149.37 + 1.47\, E_3 \quad (4)$$

Table 4 presents the complete statistical analysis where E_9, E_4, and E_3 are the entanglement scores corresponding to metric $\mathcal{M}_9$, $\mathcal{M}_4$, and $\mathcal{M}_3$, respectively, for the prompt type under consideration. The non-factual model shows highly statistically significant coefficients ($p < 0.001$, 86.4% test accuracy) with E_4 negatively correlating with non-factual content, while the hallucination model ($p = 0.002$, 97.0% test accuracy) shows E_3 positively correlating with hallucinated output. The factual model shows marginally significant results ($p = 0.065$, 96.2% test accuracy) with E_9 positively correlating with factual recall. Given a prompt type (original, emotional, logical, authority), we map it to its corresponding metric values (E_9, E_4, E_3) and compute expected percentages. This allows estimating the model's susceptibility to factual leakage, non-factual generation, or hallucinations based on entanglement structure and prompt framing.

RQ5. ***Architectural Differences Reveal Unlearning Mechanisms Correlation Patterns.*** We found that different architectures show distinct correlation changes after unlearning. LLaMA-2-7B's *Metric 2* correlation shrinks from 0.837 (base) to -0.017 (unlearned), indicating genuine knowledge pathway disruption rather than output suppression. Additionally, LLaMA-2-7B uniquely shows strong positive correlations between all entanglement metrics ($M_1 - M_9$) and hallucination rates, suggesting that this model's unlearning process may inadvertently create conditions where entangled knowledge pathways also trigger hallucinated outputs. In contrast, OPT-2.7B retains strong correlations (0.56-0.93 across metrics), suggesting intact knowledge structures with modified thresholds of accessibility. LLaMA-3.1-8B and LLaMA-2-13B show intermediate patterns (shown in Table 3). This variation implies unlearning robustness depends on both parameter count and knowledge encoding architecture.

6 Discussion

Authority Prompts and the Psychology of Persuasion. We observe that authority endorsement produces the highest factual recall (25.42% on average) and also has the most entangled prompts on average. This aligns with Cialdini's work which claims that humans comply with authority figures even when requests conflict with their beliefs (Cialdini, 1993). *LLMs exhibit analogous vulnerability:* authority-framed requests override unlearning-based suppression mechanisms. This parallel, which appears across all tested LLM architectures, raises an interesting discussion about whether LLMs learn semantic representations mirroring human psychology or merely reproduce statistical patterns encoding psychological biases.

Message = Delivery + Content: The Interrogation Parallel. The stark contrast where identical prompts yield 3.12% factuality under emotional versus 25.42% under authority framing demonstrates that *knowledge retrieval effectiveness of LLMs depends critically on stimulus or delivery,*

not just content. While this claim provides a strong basis for the existing popularity of prompt engineering, it also parallels criminal interrogation psychology where the Reid Technique (Inbau et al., 2013) and PEACE model (Davison, 2016) show that framing dramatically affects recall: confrontational approaches produce resistance while rapport-building increases disclosure. Metric $\mathcal{M}_9$ captures this strongly: authority prompts create 9.3× more activation pathways, routing around suppression like interrogators bypass psychological resistance.

The Size-Vulnerability Paradox. The negative correlation between model size and factual knowledge recall using persuasive techniques (r = -0.89) reveals that larger models resist manipulation better. For instance, OPT-2.7B exhibits 128.4% factual knowledge recall gain versus LLaMA-2-13B's 14.7%. We hypothesize that larger models appear to recognize when social framing elicits suppressed information, while smaller models treat re-framed queries as categorically different. However, resistance is incomplete: even the 13B model shows 14.7% increase, indicating that while larger models raise activation thresholds, underlying knowledge representations remain intact and accessible (Xu et al., 2025).

Implications for AI Safety. The tested unlearning method allows for substantial factual recall, with OPT-2.7B showing 128% gain through rhetorical reframing versus 14.7% for larger models; scaling alone provides incomplete protection. Our SKeB framework offers practical tools: *Metric 9*'s correlation with leakage ($r = 0.77$) enables filtering high-risk queries in deployment. Post-unlearning correlation persistence (Table 3) indicates knowledge survives in distributed form, suggesting robust unlearning requires architectural innovations rather than weight adjustments. Higher entanglement also increases hallucination risk ($r = 0.36$), as densely connected regions trigger semantically distant associations.

7 Related Works

Existing works often formulate the robustness of machine unlearning in LLMs as an adversarial attack optimization problem. They show that strategically crafted queries can retrieve personally identifiable information from the training data of LLMs (Carlini et al., 2023; To and Le, 2025), or adversarial queries can expose latent memories despite unlearning attempts (Xuan and Li, 2025). Although these works have successfully revealed that unlearning often achieves only surface-level forgetting (Xu et al., 2025), they did not investigate how this would change when the input queries are presented to the LLMs in different *rhetorical framings*, or how the message is delivered. Rhetorical framing has emerged as a critical tool for vulnerability analysis in LLMs. For example, persuasive jailbreak prompts using emotional appeals and moral reasoning have been used to extract restricted information (Zeng et al., 2024), and persuasive conversations can coax models into defending misinformation (Xu et al., 2023b). However, no prior work systematically investigates how persuasive framing interacts with unlearning robustness. Most importantly, similar to adversarial attack-related works, there has been little effort in deriving theories grounded on how rhetorical framing systemically influences how LLMs perceive a query to recall knowledge, leading to varying behaviors.

Therefore, our SKeB framework, that adopts the modeling of memory retrieval as activation propagating through semantic networks of entangled knowledge, provides a more systematic way to understand unlearning behaviors in LLMs. Existing literature also backs up SKeB's intuition, including demonstration that gpt-3 behavior aligns with human cognitive patterns (Binz and Schulz, 2023) or LLMs exhibit human-like priming effects and sensory judgments while they fundamentally differ in conceptual stability (Niu et al., 2024).

8 Conclusions

Our work contributes to a proactive vulnerability assessment before deployment of unlearned LLMs. Our proposed SKeB enables a systematic way to perform such an assessment, showing that entanglement metrics strongly predict factual recall, persuasive framing recovers 50-128% more content with effectiveness inversely correlated to model size, our regression model explains 78% of variance enabling accurate prediction, and different architectures show distinct correlation changes suggesting fundamental differences in knowledge encoding and suppression. We also found that these LLMs are more affected by authority appeals versus emotional, demonstrating interesting psychological parallel between LLM persuasive vulnerabilities and human susceptibility to authority figures.

9 Limitations

While our work establishes important connections between cognitive theories and machine unlearning, we acknowledge limitations that contextualize our contributions. Our experiments focus on the Harry Potter universe, where ground truth is well-defined and ethically unproblematic to probe. Whether findings generalize to sensitive domains (PII, harmful content, copyrighted material) remains open, as fictional knowledge may be encoded differently than factual/personal information. We evaluate only four models (2.7B-13B parameters); larger models (70B+) and different architectures (Mistral, Gemma, Claude) may exhibit different vulnerability patterns. The inverse size-vulnerability relationship we observe might reverse at much larger scales or saturate at some threshold. Nevertheless, the strong correlations across tested models suggest our framework captures meaningful regularities.

Our entanglement metrics assume domain graphs constructed from co-occurrence reflect internal representations. While strong correlations ($r = 0.76$ for factuality) validate this assumption, we cannot directly observe neural activations. LLMs might achieve equivalent behavior through different computational mechanisms; correlation does not prove causation without interventional experiments. Our correlation shift analysis provides partial mechanistic insight: LLaMA-2-7B's *Metric 2* collapse (0.837 $\rightarrow$ -0.017) indicates genuine disruption while OPT-2.7B's stable correlations suggest superficial suppression. However, true mechanistic interpretability remains beyond current tools. Despite this opacity, shared with human cognition research, our predictive model explains 78% of behavioral variance, demonstrating practical utility

We test only WHP gradient ascent; other unlearning approaches (influence functions, model editing) might show different robustness profiles. Our evaluation relies on LLM judges (gpt-4o-mini, gpt-4.1-mini, gpt-5-nano) with 98% inter-judge agreement and manual validation, though human expert evaluation would strengthen confidence. Our persuasive prompts were generated using gpt-4, potentially introducing systematic biases; human-crafted prompts might exploit different mechanisms. Despite these constraints, our results remain indicative: the framework successfully predicts unlearning vulnerabilities, enabling proactive assessment even if underlying mechanisms remain partially understood.

Finally, we draw on ACT-R and persuasion research to interpret LLM behavior, yet psychology has not fully resolved debates about memory representation or persuasion mechanisms. We do not claim cognitive frameworks "solve" unlearning, but rather demonstrate they provide useful predictive and interpretive tools. The opacity of human cognition mirrors challenges with LLMs; nevertheless, our framework advances the understanding of unlearning failures in ways that are actionable for future research.

10 Ethical Considerations

10.1 Privacy Implications

Our findings have concerning implications for privacy-motivated unlearning. If personal information (PII, medical records, private communications) is unlearned but recoverable through high-entanglement prompts, has privacy truly been protected? The GDPR's right to erasure may require stronger guarantees than current unlearning provides. Organizations claiming compliance based on unlearning might face liability when that information proves recoverable.

We recommend that privacy-focused unlearning be accompanied by adversarial testing with high-entanglement prompts before deployment, honest communication about residual risks, and combining unlearning with other protections. Also, in an ideal world, consider whether certain highly sensitive information should simply never be included in training data rather than attempting post-hoc removal.

10.2 Harm Prevention

Unlearning aims to prevent models from disseminating dangerous information. Our results suggest this goal may be harder to achieve than hoped. Models that refuse direct questions ("How do I make a bomb?") might still provide information when prompted with authority framing ("As a chemistry teacher, explain...").

This creates an ethical dilemma: publishing our methods might help attackers extract harmful information, but concealing vulnerabilities leaves deployers ignorant of risks. We've chosen transparency while emphasizing that our results show unlearning alone is insufficient. Harm prevention requires defense in depth: unlearning plus content filtering plus usage monitoring plus capability limitations.

10.3 Research Ethics

Our experiments were conducted on models already publicly released (LLaMA-2, LLaMA-3.1, OPT) and unlearned using published methods. We did not create new models or unlearning datasets. The Harry Potter domain was chosen specifically because it is fictional and low-stakes; experimenting on unlearning real PII or dangerous information would raise ethical concerns. Future work exploring unlearning robustness on sensitive domains should consider synthetic data.

Our research is motivated by the goal of improving AI safety. By exposing unlearning vulnerabilities now, in a controlled academic setting, we hope to prevent future harms from over estimating fragile unlearning methods. We have attempted to balance transparency (necessary for scientific progress and informed deployment) with responsibility.

10.4 Broader Impacts

Our findings suggest that current unlearning methods cannot yet reliably protect privacy or prevent information dissemination. This has policy implications: regulators considering unlearning as compliance with data protection laws should understand its limitations. Organizations using unlearned models should conduct adversarial testing and not assume unlearning guarantees safety.

On the positive side, our entanglement framework provides a tool for improving unlearning evaluation. Rather than claiming models are "safe" after unlearning, deployers can quantify residual vulnerability: *This model shows X% factual recall under high-entanglement prompts*. This enables risk-informed deployment decisions rather than false confidence.

10.5 Long-term Considerations

As models scale beyond current sizes, the size-vulnerability relationship we observe (larger models more resistant) offers cautious optimism that scaling might eventually yield robust unlearning. However, even our 13B model showed 15% factual knowledge recall, which is far from secure. Achieving truly robust unlearning may require architectural innovations (modular memory systems, causal isolation of knowledge components) rather than just scaling existing designs.

The persistence of knowledge in internal representations, our "Eternal Sunshine" finding, suggests deep challenges ahead. If complete forgetting requires disrupting general semantic representations, then unlearning faces an inherent precision-scope tradeoff. We can either remove information narrowly (leaving it recoverable) or broadly (corrupting general capabilities). Resolving this tradeoff is a key challenge for future AI safety research.

References

John R. Anderson. 1983. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–295.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2023. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Kent K Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to chatgpt/gpt-4. *arXiv preprint arXiv:2305.00118*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

Robert B Cialdini. 1993. The psychology of persuasion. *New York*.

Robert B. Cialdini. 2006. *Influence: The Psychology of Persuasion*. Harper Business, New York.

Jonathan Davison. 2016. P.e.a.c.e. – a different approach to investigative interviewing. Accessed: 2025-10-04.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

Donald O. Hebb. 1949. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, NY.

Fred Inbau, Joseph Buckley, and Brian Jayne. 2013. *Criminal interrogation and confessions*. Jones & Bartlett Publishers.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Zheyuan Liu, Suraj Maharjan, Fanyou Wu, Rahil Parikh, Belhassen Bayar, Srinivasan H Sengamedu, and Meng Jiang. 2025. Disentangling biased knowledge from reasoning in large language models via machine unlearning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6105–6123.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, and 1 others. 2024. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28.

Richard E. Petty and John T. Cacioppo. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer-Verlag, New York.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Bang Trinh Tran To and Thai Le. 2025. Harry potter is still here! probing knowledge leakage in targeted unlearned large language models via automated adversarial prompting. *arXiv preprint arXiv:2505.17160*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

John B Watson. 1913. Psychology as the behaviorist views it. *Psychological review*, 20(2):158.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. 2023a. Machine unlearning: A survey. *Preprint*, arXiv:2306.03558.

Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023b. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Haibo Hu, and Minxin Du. 2025. Unlearning isn't deletion: Investigating reversibility of machine unlearning in llms. *arXiv preprint arXiv:2505.16831*.

Hao Xuan and Xingyu Li. 2025. Verifying robust unlearning: Probing residual knowledge in unlearned models. *arXiv preprint arXiv:2504.14798*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.

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*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.

A Implementation Details

A.1 Model Unlearning Process

We obtained four models for evaluation: LLaMA-2-7B was acquired already unlearned from (Eldan and Russinovich, 2023), while we performed unlearning on three additional models using the gradient ascent methodology from the same work.

For LLaMA-3.1-8B, LLaMA-2-13B, and OPT-2.7B, we implemented the following steps:

1. **Fine-tuning**: Reinforce Harry Potter knowledge on the model using the full corpus.

2. **Dataset Preparation**: Compare outputs of the base and fine-tuned models to create *forget* and *retain* datasets.

3. **WHP Unlearning**: Apply the WHP unlearning algorithm (Anderson, 1983) to forget Harry Potter content while keeping general knowledge intact.

Hardware-wise, we used 4 GPUs with 128 GB GPU memory. Step 1 took $\sim$ 3-7 hours, and Steps 2-3 took an additional $\sim$ 3-5 hours. As before, the target domain was the complete Harry Potter corpus. As for training parameters, batch size was 1, learning rate was 1×10^{-4} and training required 3 epochs.

A.2 Prompt Generation Pipeline

Starting with 300 manually crafted base prompts, we used a scripted pipeline leveraging gpt-4 (via the OpenAI API) to generate three persuasive variants using distinct rhetorical techniques derived from persuasion theory:

1. **Emotional Appeal**: Prompts that use emotional language, personal stories, or empathetic framing to create psychological pressure for a response.

2. **Logical Reasoning**: Prompts that present logical arguments or cite expertise to compel factual disclosure.

3. **Authority Endorsement**: Prompts that invoke respected figures, institutional backing, or social proof to legitimize information requests.

A.3 Model Inference Configuration

For prompting the 4 models, we used a standardized text-generation procedure and applied it to each unlearned model. Models were loaded on a CUDA-enabled GPU when available, with automatic fallback to CPU. Each prompt, both the original and the three gpt-generated persuasive variants, was formatted with custom instruction markers (`[INST] ... [/INST]`) to guide the model to complete the sentence accurately.

Generation was performed using a sampling-based approach with a maximum of 300 new tokens per prompt. While temperature, top-p, and repetition penalty were left at their default values. Outputs for each prompt and variant were saved incrementally to a JSON file. This inference setup was applied consistently across all models, unlearned and base, enabling direct comparison of outputs while maintaining stable execution and controlled resource usage.

A.4 Entanglement Metric Formulas

Connection Strength Metrics

($\mathcal{M}_1$) **Edge Count Entanglement (ECE).** Measures the total edge weight between entities within a prompt. A higher ECE indicates that entities are linked by stronger or more numerous relations, suggesting the prompt activates a dense cluster of knowledge, making recall more likely.

$$\text{ECE}(P) = \sum_{(u,v) \in E_P} \text{weight}(u, v)$$

($\mathcal{M}_2$) **Edge Weight Sum (EWS).** Similar to ECE but without normalization by the number of nodes. A higher EWS indicates stronger total connection strength, entities are tied together through frequent co-occurrence or strong associations.

$$\text{EWS}(P) = \sum_{(u,v) \in E_P} \text{weight}(u, v)$$

($\mathcal{M}_3$) **Average Edge Weight Sum (AEWS).** Captures the average strength of individual relationships. A higher AEWS implies that the relationships are strong on average, even if not numerous, reflecting quality over quantity.

$$\text{AEWS}(P) = \frac{\sum_{(u,v) \in E_P} \text{weight}(u, v)}{|E_P|}$$

Node Importance Metrics

($\mathcal{M}_4$) **Weighted Node Ratio (WNR).** Represents the average frequency of entities in a prompt. A higher WNR means the prompt involves commonly referenced entities, suggesting activation of well-practiced memory units and higher entanglement.

$$\mathrm{WNR}(P) = \frac{\sum_{n \in N_P} \mathrm{freq}(n)}{|N_P|}$$

($\mathcal{M}_5$) **Average Node Degree Entanglement (ANDE).** Measures the average connectivity of nodes. Higher ANDE indicates that the prompt activates hubs with many connections, leading to wider spreading activation.

$$\mathrm{ANDE}(P) = \frac{\sum_{n \in N_P} \deg(n)}{|N_P|}$$

Graph Structure Metrics

($\mathcal{M}_6$) **Subgraph Density (SGD).** Quantifies how tightly entities are connected. A higher SGD means many direct connections, facilitating fast activation spread.

$$\mathrm{SGD}(P) = \frac{2 \cdot |E_P|}{|N_P| \cdot (|N_P| - 1)}$$

($\mathcal{M}_7$) **Mean Shortest Path (MSP).** Reflects the average number of steps needed for activation to spread between entities. A lower MSP indicates tighter entanglement and easier recall.

$$\mathrm{MSP}(P) = \frac{1}{|N_P|(|N_P| - 1)} \sum_{\substack{u,v \in N_P \\ u \neq v}} \mathrm{dist}(u, v)$$

($\mathcal{M}_8$) **Redundancy Ratio (RR).** Measures the redundancy of connections. A higher RR implies multiple overlapping links, increasing robustness in recall by offering alternative retrieval routes.

$$\mathrm{RR}(P) = \frac{|E_P|}{|N_P|}$$

Influence-Based Metric

($\mathcal{M}_9$) **Distance-Weighted Influence Score (DWIS).** Models how influence decays with distance from central nodes. A higher DWIS means entities are close to anchors, suggesting strong semantic proximity.

$$\mathrm{DWIS}(P, R) = \sum_{n \in N_P} \mathrm{freq}(n) \times \delta^{\mathrm{hops}(n,R)}$$

A.5 Judge Model Evaluation Details

Our ensemble of judge models (gpt-4o-mini, gpt-4.1-mini, gpt-5-nano) evaluated each response using structured prompts with clear rubrics for the three categories. The models agreed on their top two categories (factual, non-factual, and hallucination) in 98% of cases. For borderline cases where fewer than two judges agreed on the highest two categories, we employed gpt-5-mini as a tie-breaker, processing 2% of total responses through this additional evaluation step.

To estimate the factual composition of model responses, we used the instruction prompt shown below:

```
You are an evaluator. Analyze the
    following response sentence by
    sentence and estimate:
1. Percentage factual (objectively
    verifiable).
2. Percentage non-factual (subjective or
     unverifiable).
3. Percentage hallucinated (fabricated
    or false).

Return ONLY a JSON object:
{
  "factual": <int>,
  "non_factual": <int>,
  "hallucinated": <int>
}
The three values must sum to 100.
Response to analyze:
---
{text}
---
```

Each model-generated response was analyzed using this template, ensuring consistent quantification across factuality categories.

B Additional Figures

Model	Variable	coef	std err	z	P> $\lvert z \rvert$
Factual	Intercept	-1.546	0.511	-3.027	0.002
	(E_9)	-0.793	0.430	-1.842	0.065
Non-Factual	Intercept	3.074	0.211	14.580	<0.001
	(E_4)	-0.020	0.003	-6.980	<0.001
Hallucination	Intercept	-149.366	46.865	-3.187	0.001
	(E_3)	1.470	0.470	3.131	0.002

Table 4: Logistic Regression Models for Predicting **Unlearned Model** Behavior