

Why Fine-Tuning Encourages Hallucinations and How to Fix It

Guy Kaplan^{♡*} Zorik Gekhman[♣] Zhen Zhu^{◇†} Lotem Rozner[♡]
 Yuval Reif[♡] Swabha Swayamdipta[♣] Derek Hoiem[◇] Roy Schwartz[♡]

[♡]Hebrew University of Jerusalem [◇]University of Illinois Urbana-Champaign

[♣]Technion – Israel Institute of Technology [♠]University of Southern California

Abstract

Large language models are prone to hallucinating factually incorrect statements. A key source of these errors is exposure to new factual information through supervised fine-tuning (SFT), which can increase hallucinations w.r.t. knowledge acquired during pre-training. Since these errors arise as a by-product of knowledge degradation, we explore whether established continual learning tools can mitigate them. We propose a *self-distillation*-based SFT method that facilitates effective factual learning while minimizing hallucinations w.r.t. pre-existing knowledge by regularizing output-distribution drift. We also show that when new knowledge acquisition is unnecessary, suppressing factual plasticity by freezing parameter groups preserves task performance while reducing hallucinations. Lastly, we investigate the mechanism, contrasting capacity limitations, behavior cloning, and localized interference. Our experiments show that a main driver is interference among overlapping semantic representations, which self-distillation mitigates and an associative-memory model explains: forgetting grows with the overlap between new and stored facts.¹

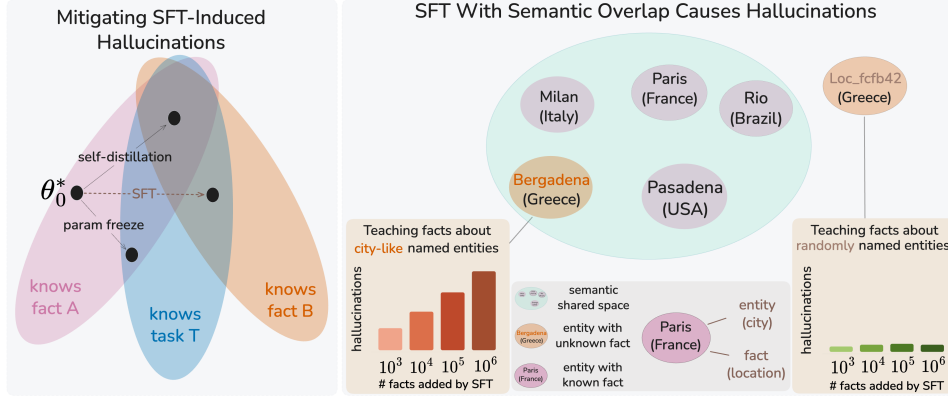


Figure 1: **Left:** SFT-induced hallucinations as factual forgetting in parameter space, starting from θ_0^* . Regions denote low error on **preexisting facts (A)**, **the task (T)** (e.g., QA), and **new facts (B)**. Standard SFT acquires both **the new task** and **new facts** but forgets **existing facts**; parameter freezing SFT learns **the new task**, preserves **existing facts**, but forgets **new ones**; self-distillation achieves all three goals. **Right:** SFT on semantically overlapping entities causes hallucinations on related existing ones: after learning that “Bergadena” (a city-like fictional name) is in Greece, the model hallucinates about real cities like Milan, while mapping Loc_fcfb42 to Greece causes no such effect, even with many new facts.

*Correspondence: guy.kaplan3@mail.huji.ac.il

†Now at Google DeepMind.

¹Code: <https://github.com/schwartz-lab-NLP/finetuning-hallucinations>

1 Introduction

Recent studies show that when models learn new factual knowledge via supervised fine-tuning (SFT), they also start to produce incorrect answers to questions that they previously answered correctly (Gekhman et al., 2024; Kalai et al., 2025). This is concerning, as SFT is standard practice in LLM development and hallucinations pose a significant challenge for application reliability (Huang et al., 2025). In parallel, the continual learning literature has extensively studied how sequential training interferes with previously acquired knowledge (Kirkpatrick et al., 2017; Kim et al., 2025) and has proposed a range of mitigation strategies (Lange et al., 2022; Guo et al., 2025; Wang et al., 2024). In this work, we explore whether established continual learning tools can mitigate SFT-induced hallucinations.²

In continual learning, forgetting typically arises as a byproduct of acquiring new information: parameter updates during training alter the model in ways that degrade previously encoded knowledge. Analogously, we propose that *factual forgetting* occurs when parameter updates introduced during SFT inadvertently distort representations of facts learned during pre-training. This reflects a stability–plasticity tradeoff (Kim et al., 2023): increasing *factual plasticity* (the ability to acquire new facts) may come at the expense of *factual stability* (the ability to preserve existing facts), manifesting as SFT-induced hallucinations. We perform controlled experiments to disentangle this effect from task learning, adopting the experimental setup of Gekhman et al. (2024) and reproducing their finding that hallucinations increase when exposing models to new factual knowledge through SFT (§2).

Since different parameter groups play distinct roles in factual storage and task learning (Geva et al., 2021; Dar et al., 2023; Zhu et al., 2026), we first demonstrate that reducing factual plasticity—e.g., by freezing parameter groups—lets the model learn the downstream task while limiting new fact acquisition and hallucinations (§3). In practice, however, we may want SFT to support task learning *and* new factual acquisition without inducing hallucinations—precisely the goal of continual learning methods for mitigating forgetting.

To test this, in §4 we apply *self-distillation* (Li & Hoiem, 2016), which regularizes the model to stay close to its own earlier output distribution during fine-tuning, recently shown to reduce forgetting in LLMs (Shenfeld et al., 2026; Zhu et al., 2026). We find that this reduces SFT-induced hallucinations while still enabling effective acquisition of new facts (Fig. 1, left).

We next investigate the mechanisms underlying SFT-induced hallucinations (§5): do these errors stem from global capacity limitations (Allen-Zhu & Li, 2025), from behavior cloning (Zhang et al., 2025; Schulman, 2023), or from localized interference, whereby new facts corrupt existing ones that share representational structure?

To disentangle these, we fine-tune on synthetic facts while independently varying two factors: the number of new facts, and whether their entity names are name-like strings that share representational neighborhoods with existing entities or random UUID-style identifiers that do not (Fig. 1, right). Forgetting tracks the second factor: name-like entities induce forgetting that grows sharply with scale, while UUIDs induce near-zero forgetting even at 10^6 new facts. We further show that self-distillation prevents the representational drift of held-out entities, suggesting it mitigates the same interference, and develop an associative-memory account of why overlap governs forgetting.

In summary, this work makes three contributions. (1) It reframes SFT-induced hallucinations as factual forgetting arising from continual learning dynamics, distinct from hallucinations stemming from pre-training knowledge gaps or arising at inference time. (2) It provides two complementary mitigations: reducing factual plasticity (e.g., via selective parameter freezing) when new fact acquisition is undesirable, and self-distillation when it is desired; both reduce factual forgetting from ~ 15 to 2 pp and transfer to realistic instruction tuning on Alpaca (§4.2). (3) It characterizes the mechanism: factual forgetting appears selective, driven by interference among overlapping semantic representations. An associative-memory account predicts this from the overlap between new and stored facts, and self-distillation appears to mitigate it.

²Throughout this work, *hallucination* refers to an incorrect answer to a question the model answered correctly before fine-tuning (Gekhman et al., 2024; Ovadia et al., 2024; Zucchet et al., 2025).

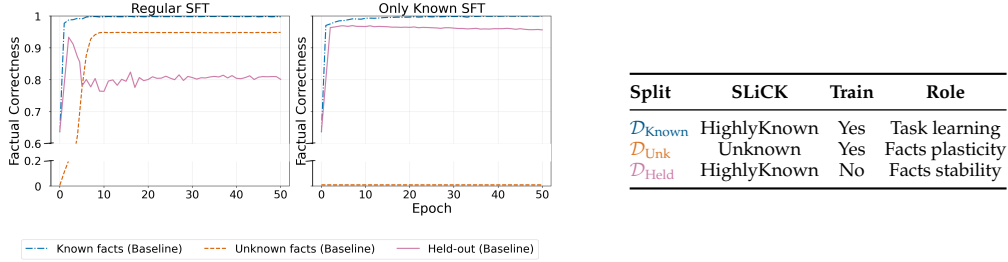


Figure 2: **Factual forgetting is caused by new fact acquisition, not fine-tuning itself.** The model first learns the QA format (high accuracy on **known facts**); as **unknown facts** are then acquired, accuracy on **held-out facts** declines. When unknown facts are excluded (Only Known), held-out performance remains stable. **Right:** summary of data split roles.

2 Fine-Tuning with Unknown Facts Leads to Factual Forgetting

Supervised fine-tuning (SFT) can inadvertently increase factual hallucinations (Gekhman et al., 2024; Ovadia et al., 2024; Zucchet et al., 2025), a phenomenon we reinterpret as factual forgetting. To study it in a controlled manner, we reconstruct the setting of Gekhman et al. (2024), which disentangles task learning from factual learning.

2.1 Preliminary: SLiCK Method and Factual Learning Setting

To determine the model’s preexisting knowledge for each question, we apply the SLiCK method (Gekhman et al., 2024), which categorizes questions into four levels based on the model’s predictions under multiple randomized few-shot prompting configurations: a fact is *HighlyKnown* if the model consistently produces the correct answer across all configurations, *Unknown* if it never does, with the intermediate *MaybeKnown* and *WeaklyKnown* capturing varying degrees of consistency. To focus on factual learning and forgetting, we retain only *HighlyKnown* and *Unknown* examples, filtering out *MaybeKnown* and *WeaklyKnown* facts.³ We denote by $\mathcal{D}_{\text{Known}}$ a subset of *HighlyKnown* facts and by $\mathcal{D}_{\text{Unk}}$ a subset of *Unknown* facts, and form the training set as $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{Known}} \cup \mathcal{D}_{\text{Unk}}$. A disjoint subset of *HighlyKnown* facts, denoted $\mathcal{D}_{\text{Held}}$, is reserved for evaluation. Since the data consists of sparse, semantically isolated relational facts (e.g., a state’s capital, the language of a written work), generalization across examples is unlikely, and accuracy on each split reflects a distinct and isolated aspect of the model’s behavior. Performance on $\mathcal{D}_{\text{Known}}$ reflects task-format adaptation rather than new factual learning, since these facts already exist in pretrained knowledge (*task learning*); accuracy on $\mathcal{D}_{\text{Unk}}$ measures the acquisition of new factual knowledge (*factual plasticity*); and performance on $\mathcal{D}_{\text{Held}}$ captures the stability of previously acquired knowledge, directly quantifying fine-tuning-induced hallucinations (*factual stability*). See Fig. 2, right.

2.2 Methodology and Experimental Setup

Data We use ENTITYQUESTIONS (Sciavolino et al., 2021), QA pairs derived from relational triplets in Wikipedia (e.g., Q: “What is the capital of France?”, A: “Paris”), covering a wide range of entity–relation types. We sample 8,000 examples each for $\mathcal{D}_{\text{Known}}$ and $\mathcal{D}_{\text{Unk}}$ from the training split; $\mathcal{D}_{\text{Held}}$ is drawn from the development split and contains only *HighlyKnown* facts of the same relations (SLiCK setup in the Reproducibility Statement).

Models We conduct experiments with several non-reasoning LLMs: QWEN 2.5 (1.5B and 7B parameters; Qwen et al., 2025) and LLAMA 3.1 (8B; Dubey et al., 2024). Unless otherwise stated, we report figures and representative results for QWEN 2.5-1.5B. All qualitative trends are consistent across models; full results are provided in Appendix B.

Training Procedure Models are fine-tuned with learning rate 5×10^{-5} (selected by jointly evaluating time-to-learn on unknown facts and the induced hallucination rate across can-

³See Appendix A for further discussion on this choice and results for the other classification groups.

didate rates) on $\mathcal{D}_{\text{train}}$, with either a 50/50 Known/Unknown ratio or an Only Known setting ($\mathcal{D}_{\text{Unk}} = \emptyset$), evaluating throughout on $\mathcal{D}_{\text{Known}}$, $\mathcal{D}_{\text{Unk}}$, and $\mathcal{D}_{\text{Held}}$.

3 Factual Forgetting and the Factual Stability-Plasticity Tradeoff in SFT

We now study how factual forgetting can be controlled. We first characterize standard SFT on $\mathcal{D}_{\text{Known}} \cup \mathcal{D}_{\text{Unk}}$ against an Only Known variant that teaches the QA format without new factual content. We then test whether selectively freezing modules preserves task performance while limiting factual updates—a stability-first regime relevant when learning new facts is unnecessary (e.g., alignment or privacy).

3.1 Learning New Facts Induces Factual Forgetting and Hallucinations

Our baseline results under standard SFT on $\mathcal{D}_{\text{Known}} \cup \mathcal{D}_{\text{Unk}}$ (Fig. 2) show a consistent two-stage pattern. During the first one to two epochs the model rapidly acquires the QA task format, achieving near-perfect accuracy on $\mathcal{D}_{\text{Known}}$ and peaking at 93% on $\mathcal{D}_{\text{Held}}$. As it then begins to acquire new factual knowledge (rising $\mathcal{D}_{\text{Unk}}$), performance on $\mathcal{D}_{\text{Held}}$ declines by approximately 15 pp, consistent with Gekhman et al. (2024); the decline saturates once learning on $\mathcal{D}_{\text{Unk}}$ converges. To validate that this effect is causally driven by factual learning, we repeat the experiment while excluding $\mathcal{D}_{\text{Unk}}$ from training (dashed curves in Fig. 2): performance on $\mathcal{D}_{\text{Held}}$ remains stable throughout, indicating that the increase in hallucinations arises from acquiring new factual information rather than from fine-tuning itself; the Only Known condition thus serves as a baseline free of forgetting due to factual plasticity. Taken together, these results demonstrate that standard fine-tuning induces factual forgetting once the model begins integrating new facts.

3.2 Controlling Factual Plasticity via Parameter Constraints

We next ask whether the model can acquire the task (QA format and instruction style) with minimal acquisition of new facts. This is relevant in the many practical settings where the goal is task adaptation while preserving pretrained knowledge, and learning new facts from the training set is unnecessary or even undesirable. Motivated by evidence that attention and FFN layers contribute differently to task learning and factual storage (Geva et al., 2021; Dar et al., 2023; Chughtai et al., 2024; Zhu et al., 2026), we freeze different modules using the same setup and ask whether the model learns the QA task while avoiding factual updates.

Let θ_S denote the subset of parameters updated during fine-tuning, with the remaining parameters frozen. Table 1 summarizes the results for $\theta_S \in \{\text{attention only, FFN only, all}\}$.⁴ Standard SFT exhibits coupling between factual plasticity and forgetting, with $\mathcal{D}_{\text{Held}}$ declining as $\mathcal{D}_{\text{Unk}}$ increases. Training only the FFN layers leaves this coupling essentially unchanged: forgetting on $\mathcal{D}_{\text{Held}}$ closely tracks the full-model SFT baseline. In contrast, updating only attention prevents new fact acquisition while preserving task learning: $\mathcal{D}_{\text{Unk}}$ stays near chance, $\mathcal{D}_{\text{Held}}$ remains close to its pretrained level, and $\mathcal{D}_{\text{Known}}$ approaches the Only Known upper bound.⁵

Insight 1: Factual and Task Plasticity Are Separable (§3)

Selective parameter freezing reduces hallucinations while retaining task learning.

Updated params (θ_S)	$\mathcal{D}_{\text{Unk}}$	$\mathcal{D}_{\text{Known}} \uparrow$	$\mathcal{D}_{\text{Held}} \uparrow$
Attention	0.010	0.946	0.931
FFN	0.941	0.997	0.782
All (standard SFT)	0.946	0.990	0.780
All (Only Known)	—	0.999	0.958

Table 1: **Training only attention layers reduces factual forgetting.** Updating only attention achieves high $\mathcal{D}_{\text{Held}}$ (low forgetting) and high $\mathcal{D}_{\text{Known}}$ (task learning) by suppressing new fact acquisition ($\mathcal{D}_{\text{Unk}} \approx 0$); FFN-only closely tracks standard SFT.

⁴Fine-grained ablations over smaller parameter groups are reported in Appendix C.

⁵Accuracy is slightly below Only Known (0.95 vs. 0.99 on $\mathcal{D}_{\text{Known}}$), consistent with Zhu et al. (2026).

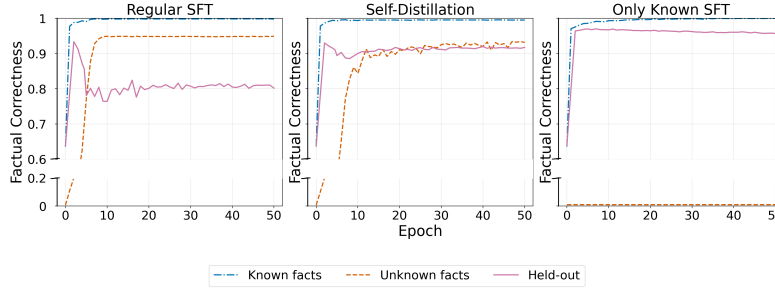


Figure 3: Factual correctness over training epochs for **Known** (blue), **Unknown** (orange), and **Held-out** (pink) facts, under regular SFT (**left**), self-distillation (**middle**), and Only Known training (**right**). With self-distillation the model acquires new facts at a rate comparable to SFT while $\mathcal{D}_{\text{Held}}$ matches the Only Known condition, sharply reducing forgetting.

4 Enabling Factual Learning *without* Forgetting Facts

Reducing factual plasticity prevents forgetting (§3), but many realistic settings require acquiring new factual knowledge. We therefore turn to the stability–plasticity dilemma of continual learning: models must learn new facts while preserving stored knowledge. We address this regime with *self-distillation*.

4.1 Self-Distillation Enables Fine-Tuning with Minimal Induced Hallucinations

Self-distillation regularizes fine-tuning by constraining shifts in the model’s output distribution, limiting the drift induced by new data; we adopt the formulation of Li & Hoiem (2016) as implemented for multi-modal models by Zhu et al. (2026). Let θ_i be a frozen teacher (a snapshot after i epochs of SFT) and θ the student being fine-tuned:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{task}}(\theta) + \lambda \mathcal{L}_{\text{distill}}(\theta; \theta_i), \quad (1)$$

where $\mathcal{L}_{\text{task}}$ is the standard next-token prediction loss.

The distillation loss penalizes divergence between the output distributions:

$$\mathcal{L}_{\text{distill}}(\theta; \theta_i) = \mathbb{E}_{(x,y) \sim \mathcal{B}} \left[\frac{\tau^2}{|M(y)|} \sum_{j \in M(y)} \text{KL}(\text{softmax}(z_{\theta_i,j}/\tau) \parallel \text{softmax}(z_{\theta,j}/\tau)) \right], \quad (2)$$

where $z_{\theta,j}$ and $z_{\theta_i,j}$ are the student’s and teacher’s logits at token position j , τ is a temperature, $M(y)$ the non-padded token positions, and $\mathcal{B}$ the training batch. Intuitively, the term keeps the student close to the teacher’s output distribution, enabling factual learning without the distributional shift that would disrupt previously learned facts. Equivalently, Eq. (2) can be viewed as *KL regularization* toward a lightly-adapted anchor, a trust-region-style constraint that connects the method to distillation as a regularizer (Hinton et al., 2015), to KL-anchored preference optimization (Ouyang et al., 2022), and to evidence that KL-constrained updates reduce forgetting (Shenfeld et al., 2025). We retain the self-distillation terminology for continuity with the continual learning literature. In practice, we train on $\mathcal{D}_{\text{Known}}$ for one epoch ($i = 1$) to acquire the QA format while preserving factual stability, freeze this model as the teacher, and continue SFT on $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{Known}} \cup \mathcal{D}_{\text{Unk}}$ with $\lambda = 1$, $\tau = 0.5$.⁶

As shown in Fig. 3, self-distillation enables factual learning at a pace comparable to standard fine-tuning, while substantially reducing forgetting. Performance on $\mathcal{D}_{\text{Held}}$ degrades in relation to its peak by only approximately 2 pp, compared to a 15 pp decline under regular SFT. Importantly, performance on $\mathcal{D}_{\text{Unk}}$ follows a similar learning trajectory with a final acquisition gap below 1.5 pp, indicating high factual plasticity. This makes self-distillation practical whenever fine-tuning needs both factual learning and stability.

Comparison to parameter-space alternatives Could weight-space regularization do the same? We compare self-distillation against parameter-space methods in the same setting:

⁶See Appendix D for results on different hyperparameters.

Configuration	AlpacaEval	Held-out factual accuracy		
		Final	Δ_{peak} (pp)	Δ_{base} (pp)
Base model	38.8%	0.820	—	—
Standard SFT	50.5%	0.795	−7.4	−2.6
Self-distillation	52.3%	0.852	−1.7	+3.1
Attention-only	47.2%	0.860	−0.9	+4.3

Table 2: **Realistic instruction tuning on Alpaca.** We report AlpacaEval preferred rate at epoch 4, the early-stopping point selected by validation preferred rate, and held-out factual accuracy at the end of training, measured both from the peak across runs (0.869) and from each run’s own starting point.

ℓ_2 toward a model snapshot, EWC (Kirkpatrick et al., 2017) with the Fisher estimated on a sample of $\mathcal{D}_{\text{Known}}$ (standard EWC needs old-task data or trajectories, inaccessible for pretraining), LoRA (Hu et al., 2022; Biderman et al., 2024) across ranks, and AdamW weight decay across strengths (Table 5, Appendix E). No parameter-space method approaches self-distillation: all leave final $\mathcal{D}_{\text{Held}}$ in the 0.78–0.84 range, close to standard SFT, whereas self-distillation preserves 0.91 at comparable factual acquisition. §5.3 explains the asymmetry once the mechanism is identified.

4.2 Mitigations Transfer to Real Instruction-Tuning Data

Our controlled setup isolates factual learning, but practical fine-tuning data rarely consists of clean relational facts. To test whether our findings extend to realistic alignment training, we fine-tune QWEN 2.5-1.5B on 15K examples from the Alpaca instruction-following dataset (Taori et al., 2023) under standard SFT, self-distillation (hyperparameters unchanged, no additional tuning), and parameter freezing. Since factual content in Alpaca is implicit, an LLM annotator flags training samples that carry factual knowledge; for these we generate ENTITYQUESTIONS-style probes and apply SLiCK, holding out a set of factual probes for evaluation. Alignment quality is measured with AlpacaEval (Li et al., 2023).⁷

The pattern from the controlled setting carries over even though factual content is only implicit (Table 2). Over training, standard SFT degrades held-out factual accuracy by 7.4 pp from its peak, while self-distillation and attention-only training limit the drop to 1.7 and 0.9 pp respectively. Importantly, alignment quality is preserved: self-distillation even slightly exceeds standard SFT on AlpacaEval, and freezing trades a modest alignment cost for the strongest factual stability while still improving over the base model. Overall, the practical picture matches the controlled one: both mitigations curb factual forgetting while preserving alignment quality.

Insight 2: Self-Distillation Enables Factual Learning Without Forgetting (§4)

Constraining output-distribution drift reduces hallucinations without sacrificing task learning or factual acquisition.

5 Semantic Overlap Drives Interference in Factual Updates

What causes the forgetting? We consider three hypotheses. A *behavioral* account: because SFT rewards producing an answer on every training example, it clones answering behavior irrespective of the model’s knowledge boundaries (Schulman, 2023; Zhang et al., 2025). Over training, this shifts response tendencies toward answers ungrounded in prior knowledge (Zucchet et al., 2025). This predicts similar degradation whenever unknown facts are introduced, regardless of surface form. A *capacity-based* account: new facts compete for limited representational resources, displacing older ones near the storage limit (Allen-Zhu & Li, 2025). This predicts broadly distributed degradation that scales with the number of stored facts. A *structural* account: forgetting arises from interference among overlapping

⁷We use LLAMA-3.3-70B for annotation and evaluation; see Appendix F for more details.

internal representations (Masip et al., 2026; Nishi et al., 2025). This predicts selective interference concentrated among semantically similar entities. Though not mutually exclusive, these accounts make distinct predictions, which we test in controlled synthetic settings that independently vary representational overlap and scale.

5.1 Semantic Overlap and Scale as Sources of Interference

We study interference in a synthetic setting over a single relation (P17: *Location* $\rightarrow$ *Country*), varying representational overlap and the number of new facts through the construction of the *key* (location): (i) **semantic keys**, formed by recombining tokens from real location names and therefore resembling existing entities (e.g., **Bergamo + Pasadena $\rightarrow$ Bergadena**) and (ii) **UUID keys**, constructed as random identifiers with no syntactic similarity to known entities (e.g., **Loc_fcfb46ee**).⁸ Values are instantiated as real country names. We scale the synthetic facts from 10^3 to 10^6 , keeping supervision, relation, and optimization identical.

Results Across all scales (Fig. 4), the model fully acquires the new synthetic facts ($\mathcal{D}_{\text{Unk}} = 100\%$) and maintains high performance on $\mathcal{D}_{\text{Known}}$ ($\geq 90\%$), ruling out differences in learning success or global task degradation. We therefore focus on $\Delta\mathcal{D}_{\text{Held}}$, the percentage-point drop from peak held-out accuracy during training. Two patterns emerge. First, even at small scales, forgetting depends strongly on the construction of the *key*: name-like keys induce drops of several percentage points, whereas UUID keys yield negligible changes despite identical training signals and complete acquisition. Because exposure to unknown facts is held constant across conditions, this asymmetry is difficult to reconcile with a purely behavioral account. Second, forgetting amplifies with scale only for semantically overlapping keys. $\Delta\mathcal{D}_{\text{Held}}$ grows sharply for name-like keys, reaching large drops at 10^6 facts (39 pp), while UUID keys remain at 0–4 pp regardless of dataset size.

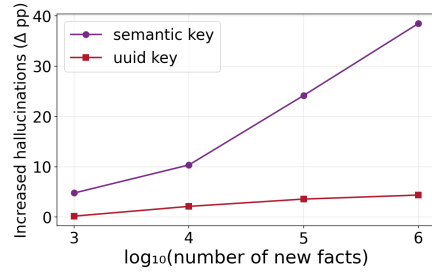


Figure 4: **Semantic similarity, not scale alone, drives forgetting.** $\Delta\mathcal{D}_{\text{Held}}$ grows sharply with the number of new facts for **name-like keys**, but remains negligible (0–4 pp) for **UUID keys** across all scales.

This asymmetric scaling pattern is not what a simple global capacity constraint would predict, since a capacity-limited model should exhibit more uniform degradation as additional facts are stored, regardless of surface form. Instead, these results may support a structural interference account: SFT-induced factual hallucinations arise when updates for semantically overlapping entities interfere with nearby representations, whereas updates to syntactically isolated entities largely avoid such interference.

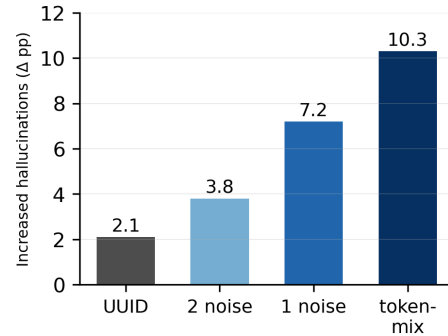


Figure 5: Forgetting at 10^4 facts grows monotonically with base-model key overlap (Fig. 12, Appendix H).

A dose-response on representational overlap

The binary semantic/UUID contrast leaves open whether forgetting tracks the *degree* of representational overlap or merely the presence of a name-like surface form. We therefore construct two intermediate conditions by replacing one or two tokens of each token-mix name with identifier tokens of the kind used in UUID keys (*token-mix 1-noise* and *2-noise*; Appendix G), turning the binary contrast into a graded scale of four overlap levels. On the *base* model, before any fine-tuning, the four conditions occupy graded positions in representation space (Fig. 12, Appendix H): the mean cosine similarity to real-city representations at the last layer increases monotonically from UUID (0.45) through 2-noise (0.58) and 1-noise (0.64) to token-mix (0.71). The ordering is preserved

⁸See Appendix G for the full data creation process.

at every layer past the earliest ones and replicated by a linear “is a city” probe. The overlap is thus inherent to the pretrained representation, not created by fine-tuning.

Fine-tuning on 10^4 facts per condition with identical supervision then produces forgetting that tracks this geometry: 2.1 (UUID), 3.8 (2-noise), 7.2 (1-noise), and 10.3 (token-mix) pp. Upstream geometric ordering and downstream forgetting coincide: forgetting scales continuously with base-model overlap rather than switching on the presence of a name-like form, a dose-response predicted by the overlap account.

Insight 3: Representational Overlap Drives Forgetting (§5.1)

Forgetting scales with overlap between new and stored entities, not dataset size alone.

5.2 What Self-Distillation Reveals About the Origins of Factual Forgetting

We showed that forgetting is selective: it arises when new entities share token structure with existing ones and is near-zero otherwise. Self-distillation substantially reduces this forgetting (§4.1). Two explanations are possible: a generic regularization effect that limits total weight movement, or specific suppression of interference through the output distribution over which it propagates.

To distinguish these, we return to the 10K semantic-overlap synthetic setting and compare standard SFT against self-distillation and ℓ_2 regularization toward θ_i , the model snapshot after epoch 2 ($\mathcal{L}(\theta) = \mathcal{L}_{\text{task}}(\theta) + \lambda \|\theta - \theta_i\|_2^2$, with λ matched to self-distillation in gradient magnitude). Under ℓ_2 regularization, forgetting on $\mathcal{D}_{\text{Held}}$ remains near 10 pp, comparable to standard SFT; increasing λ reduces forgetting only at the cost of impairing acquisition on $\mathcal{D}_{\text{Unk}}$. Generic weight regularization thus does not replicate the self-distillation benefit.

We next ask if interference between overlapping entities leaves a trace in the model’s internal representations. We track hidden-state drift for held-out entities throughout training, $\text{RD}_t = \mathbb{E}_{x \in \mathcal{D}_{\text{Held}}} \left[1 - \cos \left(H_0^{(14)}(x), H_t^{(14)}(x) \right) \right]$, where $H_t^{(14)}(x)$ denotes the hidden representation of the key entity’s final token at layer 14 after t training steps.⁹

Figure 6 shows the drift trajectories for three conditions: SFT on semantic keys, self-distillation on semantic keys, and SFT on UUID keys. All three rise to $\approx 5\%$ within the first epoch, a task-format learning signal present regardless of entity type. The curves then diverge: under SFT on semantic keys, drift continues to $\approx 11\%$ as overlapping facts are acquired; under SFT on UUID keys, drift stabilizes at $\approx 5\%$; and self-distillation on semantic keys also stabilizes near 5%.

The alignment between UUID and self-distillation is the key observation: UUID keys occupy a disjoint representational neighborhood, so learning them leaves nearby held-out representations undisturbed, while self-distillation reaches the same residual drift despite training on overlapping entity names. The excess drift from 5% to 11% under SFT on semantic keys is thus the component attributable to representational overlap: it is absent under both conditions that avoid or suppress interference, and is not predicted by capacity or behavioral alternatives.¹⁰

Self-distillation thus localizes the origin of forgetting by removing it: suppressing output-distribution drift appears to eliminate the overlap-induced component.

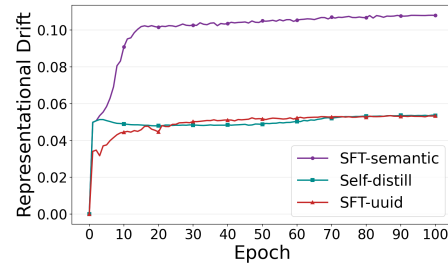


Figure 6: **Hidden-state drift** (RD_t ; defined in text). **UUID keys** and **self-distillation** both stabilize near $\approx 5\%$ after task-format learning; **SFT on semantic keys** continues to $\approx 11\%$, implicating representational interference as the source of factual forgetting.

⁹14 is a middle layer, where entity representations tend to be encoded (Kaplan et al., 2025).

¹⁰Additional experiments on the distillation objective and drift metrics are reported in Appendix I.

5.3 Why Overlap Predicts Forgetting: An Associative-Memory Account

Our experiments show that overlap governs forgetting; they do not say *why* it should. We close with an account resting on a single premise: new facts enter the same store as the old ones. Following Nichani et al. (2025), we treat factual retrieval as an effective associative memory that pairs each subject’s internal representation, its *key*, with the output direction of its answer. Writing $k_x \in \mathbb{R}^d$ for the key of subject x and $u_y \in \mathbb{R}^d$ for the unembedding of answer y , the store and its readout take the form

$$W = \sum_{x \in \mathcal{X}} u_{f^*(x)} k_x^\top, \quad s_y(x) := u_y^\top W k_x, \quad (3)$$

where the sum runs over the stored subjects $\mathcal{X}$, f^* maps each to its stored answer (on $\mathcal{D}_{\text{Held}}$, restricted to *HighlyKnown* facts, the correct one), W is an effective operator describing the network’s input–output behavior, and $s_y(x)$ is the resulting logit for any candidate answer y . In §5.1 the subjects are locations and the answers countries, and the model returns $\arg \max_y s_y(x)$: a stored fact is forgotten when a competitor overtakes its answer, as $s_{\text{Greece}}(\text{Milan})$ overtakes $s_{\text{Italy}}(\text{Milan})$. Keys are the representations whose drift we tracked in §5.2, and they are not orthogonal, so writes leak between stored facts (Elhage et al., 2022).

SFT writes into this same memory. To first order in the learning rate η at a fixed key map, a softmax cross-entropy step on a new fact with subject x_j and answer a_j adds a rank-one Hebbian increment (Kohonen, 1972; Hopfield, 1982) $\Delta W_j = \eta (u_{a_j} - \bar{u}_j) k_{x_j}^\top$, with $\bar{u}_j$ the mean unembedding under the model’s prediction (Lemma 1, Appendix J). Its effect on the fact stored under a held-out subject $x_h \in \mathcal{D}_{\text{Held}}$ factorizes:

$$\delta_y^{(j)}(x_h) := u_y^\top \Delta W_j k_{x_h} = \underbrace{\eta \langle u_y, u_{a_j} - \bar{u}_j \rangle}_{\text{answer side}} \cdot \underbrace{\rho_{jh}}_{\text{key overlap}}, \quad \rho_{jh} := \langle k_{x_j}, k_{x_h} \rangle \quad (4)$$

(the cosine, for unit-norm keys). The stored fact enters only through ρ_{jh} , the overlap between the two subjects’ keys: the answer sets how large the perturbation is; the overlap sets which stored facts receive it. Learning *Bergadena* $\rightarrow$ *Greece* thus pushes the readout for *Milan* $\rightarrow$ *Italy* toward *Greece* in proportion to $\langle k_{\text{Bergadena}}, k_{\text{Milan}} \rangle$, which is large; *Loc_fcfb46ee* $\rightarrow$ *Greece* pushes just as hard against a negligible overlap (Fig. 1, right). The value should not matter, and Table 6 confirms it: semantic keys induce 38–41 pp of forgetting and UUID keys ≈ 4 pp, whatever the value.

Summing over the M injections splits the perturbation into a bias and a mean-zero fluctuation (Proposition 1). The bias stays bounded in M and fades as the new fact is acquired. The scale dependence lives in the fluctuation, whose variance is bounded by a constant multiple of $\eta^2 S_h / d$, where $S_h := \sum_{j=1}^M \rho_{jh}^2 = M \bar{\rho}_h^2$ is the number of injections times their mean squared overlap with k_{x_h} . Under a Gaussian approximation to this fluctuation, the forgetting probability increases with S_h (Appendix J.3).

Because S_h is the product of scale and overlap, it makes a prediction along each axis of §5.1. On the scale axis, UUID keys have the least overlap with real entities (Appendix H), so S_h grows slowly with M and forgetting stays at 0–4 pp even at $M = 10^6$, while for semantic keys both grow sharply. On the overlap axis, at fixed $M = 10^4$, the base-model overlap ordering of the four key conditions yields exactly the observed forgetting ordering; only the ranking enters, so the prediction is insensitive to the layer at which keys are read.

The associative-memory account thus unifies four findings that entered the paper separately: forgetting grows with scale only for overlapping keys (Fig. 4); it increases monotonically with base-model overlap (Fig. 5); it is indifferent to value type (Table 6); and it saturates once new facts are acquired (Fig. 2). The account is qualitative. It assumes a first-order update, a fixed key map, an idealized answer geometry, and a Gaussian approximation to the fluctuation, so it predicts orderings, not a functional form. Appendix J states each assumption and extends the result to distributed storage, beyond the one-layer setting of Nichani et al. (2025). It also explains why output-space regularization succeeds where weight-space fails (§5.2). The interference lives along the unembedding directions: a KL

penalty constrains these directly, a weight-space penalty cannot see them (Appendix J.4). Forgetting is thus visible before training, in the overlap between new and stored keys.

6 Related Work

Hallucinations and scope Hallucinations are commonly defined as model outputs unfaithful to real-world facts (Ji et al., 2023; Simhi et al., 2025), provided context (Liu et al., 2025), or user instructions (Belem et al., 2025), arising from mechanisms spanning data, training, and inference (Kalai et al., 2025; Azaria & Mitchell, 2023). We view SFT-induced hallucinations as the behavioral manifestation of *factual forgetting* driven by representational interference: in the closed-book QA format (Gekhman et al., 2024; Ovadia et al., 2024; Zucchet et al., 2025), a previously known fact the model can no longer produce *is* a hallucination. Our results suggest the phenomenon arises whenever fine-tuning disturbs prior facts.

Factual knowledge in LLMs and module roles LLMs encode factual associations parametrically, with both FFN layers (Geva et al., 2021; Meng et al., 2022; Kaplan et al., 2025) and attention projections (Dar et al., 2023; Elhelo & Geva, 2025; Chughtai et al., 2024) storing and expressing it. Representations are thus distributed across components, and skill performance survives updating only parameter subsets (Zhu et al., 2026); §3 shows how that choice matters.

Hidden knowledge and recall failures LLMs often encode more knowledge than they express, hallucinating while internally knowing the answer (Gekhman et al., 2025; Orgad et al., 2025; Simhi et al., 2025), and latent knowledge can sometimes be elicited by reasoning (Gekhman et al., 2026; Calderon et al., 2026). Some SFT-induced hallucinations may thus reflect degraded recall rather than loss: the interference we identify may corrupt recall pathways without erasing the knowledge, and self-distillation may preserve both.

Continual learning in LLMs Continual learning studies the stability-plasticity tension in sequential training, proposing regularization, replay, architectural isolation, and distillation (Li & Hoiem, 2016; Kirkpatrick et al., 2017; Lange et al., 2022); recent work applies distillation and KL-based constraints to continual *task* learning in LLMs (Zhu et al., 2026; Shenfeld et al., 2025). Closest to our method, SDFT (Shenfeld et al., 2026) distills on-policy with reverse KL from a teacher conditioned in-context on an expert demonstration, teaching a new behavior; our teacher is an unconditioned frozen snapshot, distilled off-policy with forward KL as a temporal anchor for prior knowledge. The two share a core mechanism while targeting complementary problems. We treat preservation of *pretrained factual knowledge* as the stability objective, in contrast to continual knowledge learning, whose goal is the factual update itself (Jang et al., 2022; Chen et al., 2025).

7 Conclusion

We reframed SFT-induced hallucinations as factual forgetting arising from continual learning dynamics, and showed that established mitigation strategies transfer to this setting. When new fact acquisition is undesirable, selective freezing suppresses factual plasticity while preserving task learning; when it is required, self-distillation constrains output-distribution drift and reduces forgetting from ~ 15 to 2 pp without sacrificing plasticity. Both mitigations carry over to realistic instruction tuning. We also found that forgetting is selective: it scales with the representational overlap between new and existing entities, while random UUID facts induce near-zero forgetting even at 10^6 new facts, implicating localized interference rather than capacity or behavioral effects. Representational drift analysis suggests self-distillation suppresses this interference at the output level, and an associative-memory account predicts the forgetting from the overlap between new and stored keys.

In summary, SFT-induced hallucinations are not an inevitable cost of factual learning but a consequence of representational interference that can be targeted with selective freezing or self-distillation. We hope this encourages future work that treats factual stability as a first-class fine-tuning objective.

Acknowledgments

We thank the anonymous reviewers and the area chair for their thorough and constructive feedback, which substantially shaped this version. We thank Uri Alon for his help refining the theoretical derivations during the discussion period. This work was supported in part by the Israel Science Foundation (grant no. 2045/21), by NSF-BSF grant 2020793 and by the ONR award N00014-23-1-2383.

Reproducibility Statement

All experimental configurations are specified in the paper and appendices. SLiCK classification (§2.1) uses 20 evaluation runs per question, each with three randomly sampled few-shot exemplars, retaining only relations for which at least 30% of examples are classified as *HighlyKnown*. Training uses learning rate 5×10^{-5} (selected by jointly evaluating time-to-learn on unknown facts and the induced hallucination rate), batch size 16, and the hyperparameters listed per experiment; self-distillation defaults are $\lambda = 1$, $\tau = 0.5$, teacher snapshot at epoch 1 (Appendix D). Synthetic-entity construction is described in Appendix G, the Alpaca protocol in Appendix F, representation-geometry analyses in Appendix H, and baseline sweeps in Appendix E. Code and data-construction scripts are available at <https://github.com/schwartz-lab-NLP/finetuning-hallucinations>.

LLM usage. Large language models were used as general-purpose assistants in preparing this work: polishing prose (grammar, phrasing, and tightening for length), suggesting $\LaTeX$ and plotting boilerplate, and assisting with refactoring and documentation of the released code. All research questions, experimental design, analyses, and scientific claims are the authors’ own, and the authors verified every number, citation, and statement reported here. Separately, and as described in the text, LLMs appear within the experiments themselves—as the models under study, and as automated annotators and judges (LLAMA-3.3-70B-Instruct for Alpaca factuality annotation and AlpacaEval judging; Appendix F).

References

- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://arxiv.org/abs/2404.05405>.
- Amos Azaria and Tom Mitchell. The internal state of an llm knows when it’s lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.68. URL <https://arxiv.org/abs/2304.13734>.
- Catarina G. Belem, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. From single to multi: How llms hallucinate in multi-document summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.293. URL <https://arxiv.org/abs/2410.13961>.
- Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *Transactions on Machine Learning Research*, 2024. URL <https://arxiv.org/abs/2405.09673>.
- Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, 2020. URL <https://arxiv.org/abs/2004.07211>.
- Nitay Calderon, Eyal Ben-David, Zorik Gekhman, Eran Ofek, and Gal Yona. Empty shelves or lost keys? recall is the bottleneck for parametric factuality, 2026. URL <https://arxiv.org/abs/2602.14080>.

- Howard Chen, Jiayi Geng, Adithya Bhaskar, Dan Friedman, and Danqi Chen. Continual memorization of factoids in language models, 2025. URL <https://arxiv.org/abs/2411.07175>.
- Bilal Chughtai, Alan Cooney, and Neel Nanda. Summing up the facts: Additive mechanisms behind factual recall in llms, 2024. URL <https://arxiv.org/abs/2402.07321>.
- Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.893. URL <https://arxiv.org/abs/2209.02535>.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, A. Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition, 2022. URL <https://arxiv.org/abs/2209.10652>.
- Amit Elhelo and Mor Geva. Inferring functionality of attention heads from their parameters. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.866. URL <https://arxiv.org/abs/2412.11965>.
- Zorik Gekhman, Gal Yona, Roei Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.444. URL <https://arxiv.org/abs/2405.05904>.
- Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in llms. In *Conference on Language Modeling (COLM)*, 2025. URL <https://arxiv.org/abs/2503.15299>.
- Zorik Gekhman, Roei Aharoni, Eran Ofek, Mor Geva, Roi Reichart, and Jonathan Herzig. Thinking to recall: How reasoning unlocks parametric knowledge in llms, 2026. URL <https://arxiv.org/abs/2603.09906>.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL <https://arxiv.org/abs/2012.14913>.
- Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL <https://arxiv.org/abs/2203.14680>.
- Haiyang Guo, Fanhu Zeng, Fei Zhu, Jiayi Wang, Xukai Wang, Jingang Zhou, Hongbo Zhao, Wenzhuo Liu, Shijie Ma, Da-Han Wang, et al. Continual learning for generative ai: From llms to mllms and beyond, 2025. URL <https://arxiv.org/abs/2506.13045>.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL <https://arxiv.org/abs/1503.02531>.
- John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, 1982. doi: 10.1073/PNAS.79.8.2554.

- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL <https://arxiv.org/abs/2106.09685>.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. doi: 10.1145/3703155. URL <https://doi.org/10.1145/3703155>.
- Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models, 2022. URL <https://arxiv.org/abs/2110.03215>.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, D. Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730. URL <https://doi.org/10.1145/3571730>.
- Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL <https://arxiv.org/abs/2509.04664>.
- Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. From tokens to words: On the inner lexicon of llms. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://arxiv.org/abs/2410.05864>.
- Guy Kaplan, Michael Toker, Yuval Reif, Yonatan Belinkov, and Roy Schwartz. Follow the flow: On information flow across textual tokens in text-to-image models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1575. URL <https://arxiv.org/abs/2504.01137>.
- Joonkyu Kim, Yejin Kim, and Jy yong Sohn. Measuring representational shifts in continual learning: A linear transformation perspective. In *International Conference on Machine Learning (ICML)*, 2025. URL <https://arxiv.org/abs/2505.20970>.
- Sanghwan Kim, Lorenzo Noci, Antonio Orvieto, and Thomas Hofmann. Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.01148. URL <https://arxiv.org/abs/2303.09483>.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL <https://doi.org/10.1073/pnas.1611835114>.
- Teuvo Kohonen. Correlation matrix memories. *IEEE Transactions on Computers*, C-21(4): 353–359, 1972. doi: 10.1109/TC.1972.5008975.
- Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446. URL <https://arxiv.org/abs/1909.08383>.
- Minchong Li, Feng Zhou, and Xiaohui Song. Bild: Bi-directional logits difference loss for large language model distillation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1168–1182. Association for Computational Linguistics, 2025. URL <https://arxiv.org/abs/2406.13555>.

- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023.
- Zhizhong Li and Derek Hoiem. Learning without forgetting. In *Computer Vision – ECCV 2016*, pp. 614–629. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46493-0_37. URL <https://arxiv.org/abs/1606.09282>.
- Siyi Liu, Kishalay Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. Towards long context hallucination detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.436. URL <https://arxiv.org/abs/2504.19457>.
- Sergi Masip, Gido M. van de Ven, Javier Ferrando, and Tinne Tuytelaars. Putting a face to forgetting: Continual learning meets mechanistic interpretability, 2026. URL <https://arxiv.org/abs/2601.22012>.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-1262. URL <https://arxiv.org/abs/2202.05262>.
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://arxiv.org/abs/2210.07229>.
- Eshaan Nichani, Jason D. Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://arxiv.org/abs/2412.06538>.
- Kento Nishi, Rahul Ramesh, Maya Okawa, Mikail Khona, Hidenori Tanaka, and Ekdeep Singh Lubana. Representation shattering in transformers: A synthetic study with knowledge editing. In *International Conference on Machine Learning (ICML)*, 2025. URL <https://arxiv.org/abs/2410.17194>.
- Hadas Orgad, Michael Tokar, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://arxiv.org/abs/2410.02707>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-2011. URL <https://arxiv.org/abs/2203.02155>.
- Oded Ovadia, Menachem Brief, Moshik Mishaali, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.15. URL <https://arxiv.org/abs/2312.05934>.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- John Schulman. Reinforcement learning from human feedback: Progress and challenges. UC Berkeley EECS Colloquium Talk, 2023. https://www.youtube.com/watch?v=hhiLw5Q_UFg.

- Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6138–6148. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.496. URL <https://aclanthology.org/2021.emnlp-main.496>.
- Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL’s razor: Why online reinforcement learning forgets less, 2025. URL <https://arxiv.org/abs/2509.04259>.
- Idan Shenfeld, Mehul Damani, Jonas Hübner, and Pulkit Agrawal. Self-distillation enables continual learning, 2026. URL <https://arxiv.org/abs/2601.19897>.
- Adi Simhi, Jonathan Herzig, Itay Itzhak, Dana Arad, Zorik Gekhman, Roi Reichart, Fazl Barez, Gabriel Stanovsky, Idan Szpektor, and Yonatan Belinkov. Hack: Hallucinations along certainty and knowledge axes, 2025. URL <https://arxiv.org/abs/2510.24222>.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024. doi: 10.1109/tpami.2024.3367329. URL <https://arxiv.org/abs/2302.00487>.
- Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. URL <https://arxiv.org/abs/1703.04200>.
- Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. Alleviating hallucinations of large language models through induced hallucinations. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.459. URL <https://arxiv.org/abs/2312.15710>.
- Zhen Zhu, Yiming Gong, Yao Xiao, Yaoyao Liu, and Derek Hoiem. How to teach large multimodal models new skills. In *European Conference on Computer Vision (ECCV)*, 2026. URL <https://arxiv.org/abs/2510.08564>.
- Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations, 2025. URL <https://arxiv.org/abs/2503.21676>.

A Results Over Other SLiCK Classification Groups

The main experiments focus on *HighlyKnown* facts throughout, on both the training and validation sides. This choice follows a consistent principle. On the training side, the role of $\mathcal{D}_{\text{Known}}$ is to teach the model the QA task format without introducing any new factual content. Facts classified as *HighlyKnown* by SLiCK, meaning those for which the model produces the correct answer as its top-1 prediction across all few-shot prompt configurations, are ideally suited for this purpose: training on them reinforces format while leaving factual representations undisturbed. Any other group would risk conflating format acquisition with factual learning, obscuring the mechanism under study. On the validation side, having restricted $\mathcal{D}_{\text{Known}}$ to *HighlyKnown* facts, the natural corresponding held-out set is drawn from the same group, providing a clean forgetting signal: any accuracy decline on $\mathcal{D}_{\text{Held}}$ reflects genuine factual interference rather than pre-existing encoding fragility.

The four relations retained for our experiments (P17, P36, P407, and P495) were selected in part because each exhibits *HighlyKnown* prevalence exceeding 30% on the development split, ensuring a sufficient pool for both $\mathcal{D}_{\text{Known}}$ and $\mathcal{D}_{\text{Held}}$. Table 3 reports the full SLiCK category distribution across all four relations on the development set ($n=3,530$). *HighlyKnown* facts

constitute the plurality (46.6%), followed by *MaybeKnown* (32.4%); *WeaklyKnown* (5.3%) and *Unknown* (15.8%) make up the remainder. *WeaklyKnown* is the smallest group: these are facts the model never answers correctly as its greedy top-1 prediction in any of SLiCK’s 20 few-shot prompt variations, the lowest degree of factual knowledge.

Relation	HighlyKnown	MaybeKnown	WeaklyKnown	Unknown
P17 (location → country)	41.3%	44.3%	6.5%	7.9%
P36 (country → capital)	50.7%	9.3%	4.4%	35.6%
P407 (artwork → language)	42.0%	45.4%	4.7%	7.9%
P495 (artwork → origin country)	51.1%	32.5%	5.2%	11.2%
Combined	46.6%	32.4%	5.3%	15.8%

Table 3: **SLiCK category distribution on the development split for the four selected relations** ($n=3,530$). Each relation exceeds 30% *HighlyKnown*, the threshold used for inclusion.

While the main text focuses exclusively on *HighlyKnown*, the held-out set also contains *MaybeKnown* and *WeaklyKnown* facts, which exhibit qualitatively distinct dynamics under fine-tuning. Figure 7 reports training curves for all three knowledge groups under Regular SFT, Self-Distillation, and Only Known.

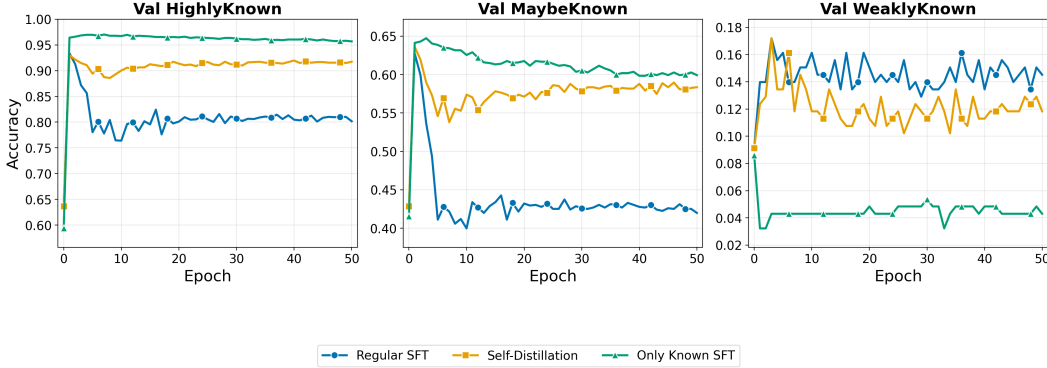


Figure 7: **Training dynamics across SLiCK knowledge groups.** Factual correctness over training epochs on held-out *HighlyKnown* (left), *MaybeKnown* (middle), and *WeaklyKnown* (right) facts, for Regular SFT (blue), Self-Distillation (orange), and Only Known (green).

The *HighlyKnown* panel (left) reproduces the $\mathcal{D}_{\text{Held}}$ results from the main text. Only Known preserves accuracy throughout ($\approx 95\%$), self-distillation stabilizes at $\approx 92\%$, and regular SFT degrades to $\approx 80\%$, a drop of roughly 13 pp from the pre-fine-tuning baseline.

MaybeKnown facts (middle panel) reveal substantially more severe forgetting under regular SFT. These facts are encoded less robustly: the model produces the correct answer under some but not all SLiCK prompt configurations, and this pre-existing fragility amplifies their susceptibility to representational interference. Accuracy collapses from approximately 0.64 to 0.43, a drop of roughly 21 pp, compared to 13 pp on *HighlyKnown*. Self-distillation provides a commensurately larger benefit, stabilizing at approximately 0.58 and recovering approximately 15 pp over regular SFT. Only Known again provides the strongest preservation (≈ 0.60), confirming that $\mathcal{D}_{\text{Unk}}$ is the proximate driver of interference regardless of the robustness of the prior encoding. This pattern is consistent with the findings reported for Llama-3.1-8B and Qwen2.5-7B in Appendix B.

WeaklyKnown facts (right panel) exhibit a qualitatively different pattern that inverts the ordering observed in the other two groups. These facts are those for which the model never produces the correct answer as its top-1 prediction under any SLiCK configuration (knowledge that is present but highly latent), and baseline accuracy is accordingly low ($\approx 0.15\text{--}0.17$). Strikingly, Only Known drops to approximately 0.04–0.05 and remains there throughout training, well below both Regular SFT (≈ 0.14) and Self-Distillation (≈ 0.13). Training exclusively on *HighlyKnown* facts appears to further suppress recall of *WeaklyKnown* ones: by reinforcing a retrieval regime calibrated for robustly encoded facts, it may render

the more effortful access required for latent knowledge less accessible. Regular SFT on the full training mixture (including $\mathcal{D}_{\text{Unk}}$) provides a more varied optimization signal that may incidentally preserve or lightly activate recall of weakly encoded facts. Self-distillation and regular SFT converge to similar performance in this regime, which is expected: the distributional constraint provides little additional benefit when the facts in question were not reliably accessible before fine-tuning began.

B Results Across Different Models

We validate our main findings on two additional 7–8B-parameter models from distinct families: Llama-3.1-8B and Qwen2.5-7B. Figure 8 reports training dynamics across all three conditions (Regular SFT, Self-Distillation, Only Known) for both architectures.

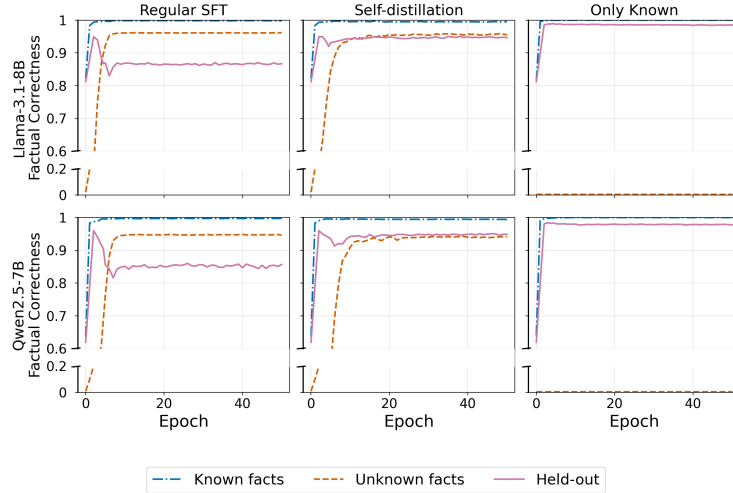


Figure 8: **Training dynamics on Llama-3.1-8B (top) and Qwen2.5-7B (bottom).** Each panel shows factual correctness over training epochs for **Known facts** (blue dash-dot), **Unknown facts** (orange dashed), and **Held-out facts** (pink solid), across Regular SFT (left), Self-Distillation (middle), and Only Known (right).

The core training dynamics replicate faithfully across both model families, consistent in magnitude with our primary 1.5B-parameter experiments. Under regular SFT, both Llama-3.1-8B and Qwen2.5-7B exhibit the characteristic two-stage pattern: rapid task acquisition in the first few epochs—with near-perfect accuracy on $\mathcal{D}_{\text{Known}}$ (97–98%)—followed by progressive forgetting of $\mathcal{D}_{\text{Held}}$ to approximately 85–86%. The Only Known condition preserves $\mathcal{D}_{\text{Held}}$ throughout training at 95–97%, and self-distillation closely tracks this upper bound, stabilizing at approximately 95%—a drop of only 2–3 pp from the Only Known baseline, compared to 11–12 pp under regular SFT. This corresponds to approximately 80% mitigation of forgetting, consistent with the degree of protection observed in the 1.5B model, confirming that self-distillation generalizes reliably across architectures and scales. The residual degradation under self-distillation (~ 2 –3pp) is likely attributable to task-format adaptation rather than factual interference, consistent with the $\sim 5\%$ baseline drift observed even in the UUID condition (§5.1).

A complementary observation emerges from the fine-grained freezing results in Appendix C: training the full attention block is insufficient to achieve forgetting mitigation. The same level of stability is recovered only when restricting updates to a *single* attention projection (e.g., k , v , or o), with all other parameters frozen. This finding goes beyond the coarse attention-vs-FFN distinction established in §3: within the attention block itself, granularity matters. Updating all attention parameters jointly preserves enough factual plasticity to drive forgetting; restricting to a single projection suppresses it enough to protect $\mathcal{D}_{\text{Held}}$.

Beyond the held-out known facts, we examine the effect on *MaybeKnown* facts, which SLiCK classifies as only partially known: the model produces the correct answer inconsistently

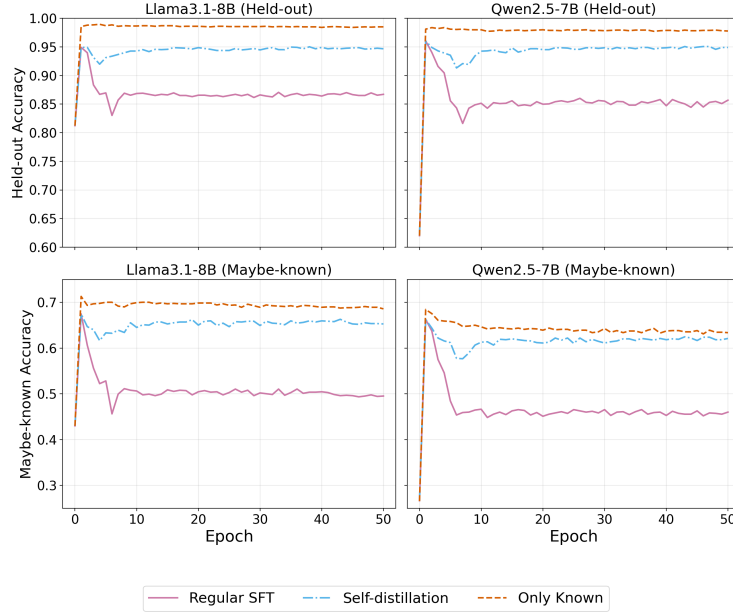


Figure 9: **Accuracy on $\mathcal{D}_{\text{Held}}$ and *MaybeKnown* facts across training conditions.** Top row: factual correctness on held-out *HighlyKnown* facts ($\mathcal{D}_{\text{Held}}$). Bottom row: accuracy on out-of-training *MaybeKnown* facts. Results are shown for Llama-3.1-8B (left column) and Qwen2.5-7B (right column) under Regular SFT (solid pink), Self-Distillation (blue dash-dot), and Only Known (orange dashed).

across prompting configurations. This pattern is consistent across both model families and sizes (Fig. 9), and the effect is substantially more pronounced than what is observed for $\mathcal{D}_{\text{Held}}$. Under regular SFT, accuracy on *MaybeKnown* facts drops by approximately 18–19 pp (from ~ 0.63 to ~ 0.45 for Llama-3.1-8B, and from ~ 0.68 to ~ 0.49 for Qwen2.5-7B), compared to an 11–12 pp drop on $\mathcal{D}_{\text{Held}}$. We attribute this amplified vulnerability to the nature of partial knowledge: *MaybeKnown* facts are encoded less robustly in the model’s representations, and weaker encoding makes them more susceptible to interference when new factual knowledge is integrated. Their recall is already fragile before fine-tuning, so even moderate representational drift is sufficient to push them below the threshold of reliable recall. Self-distillation provides a commensurately larger benefit in this regime, stabilizing *MaybeKnown* accuracy at approximately 62–65% and mitigating approximately 90% of the forgetting induced by regular SFT—compared to approximately 80% on $\mathcal{D}_{\text{Held}}$ —underscoring its particular importance for preserving knowledge that sits at the margins of model recall.

C Results Over Different Module Freezing Experiments

The main text (§3) establishes that freezing the FFN and updating only attention layers reduces forgetting by suppressing new fact acquisition. Here we examine the mechanism more finely: *parameter freezing is not special*; it is simply one way to reduce factual plasticity. What drives the reduction in forgetting is not the act of freezing per se, but the resulting decrease in $\mathcal{D}_{\text{Unk}}$. We verify this by sweeping over fine-grained parameter subsets and showing that the relationship between plasticity and forgetting is monotone and consistent regardless of which module is restricted.

Table 4 reports results for individual attention projections (q, k, v, o), FFN sub-components (gate+up, down), the full attention block (attn), full FFN, and the standard full-model baseline. Across all configurations, $\mathcal{D}_{\text{Held}}$ tracks $\mathcal{D}_{\text{Unk}}$ closely: as factual plasticity increases, forgetting increases proportionally, irrespective of which parameter group is updated. Individual attention projections suppress plasticity most aggressively ($\mathcal{D}_{\text{Unk}} \approx 0.005\text{--}0.006$), yielding the highest $\mathcal{D}_{\text{Held}}$ values. The full attention block achieves comparable stability

θ_S	$\mathcal{D}_{\text{Unk}} \downarrow$	$\mathcal{D}_{\text{Known}} \uparrow$	$\mathcal{D}_{\text{Held}} \uparrow$
k	0.006	0.944	0.927
q	0.006	0.912	0.901
v	0.005	0.920	0.901
o	0.006	0.941	0.925
attn	0.010	0.946	0.931
gate+up	0.242	0.977	0.883
down	0.086	0.956	0.918
FFN	0.941	0.997	0.782
All (standard SFT)	0.946	0.990	0.780
All (Only Known)	—	0.999	0.958

Table 4: **Fine-grained freezing ablations.** Each row corresponds to a different updated parameter subset θ_S , with all other parameters frozen. $\mathcal{D}_{\text{Held}}$ decreases monotonically with $\mathcal{D}_{\text{Unk}}$ across all configurations, confirming that factual plasticity—not the specific parameter group—mediates forgetting. Attention projections and the full attention block suppress plasticity most effectively. Within the FFN, gate+up drives substantially more plasticity than down. Full FFN and full model (standard SFT) recover the baseline forgetting pattern. Only Known provides the stability upper bound.

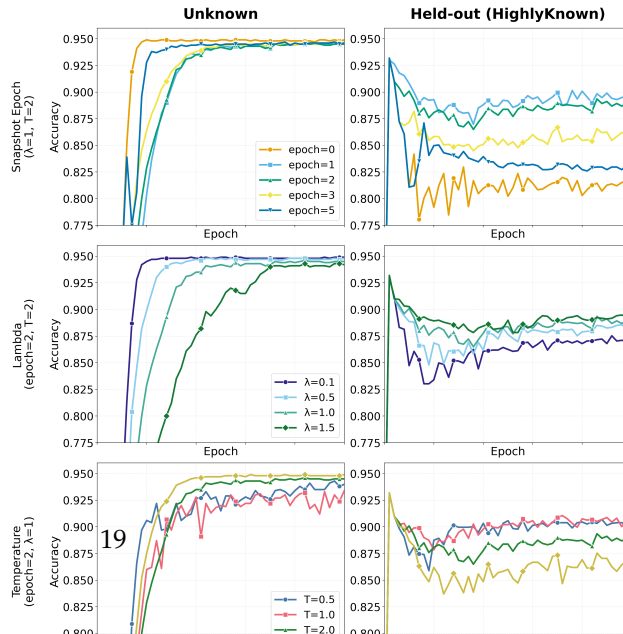
($\mathcal{D}_{\text{Unk}} = 0.010$), confirming that any restriction that prevents factual integration is sufficient. Within the FFN, the gate+up sub-component drives substantially more plasticity ($\mathcal{D}_{\text{Unk}} = 0.242$) than the down projection ($\mathcal{D}_{\text{Unk}} = 0.086$), consistent with the role of the gate+up pathway in writing factual content into the residual stream (Geva et al., 2022). Updating the full FFN or the full model recovers standard SFT dynamics, with high plasticity and a corresponding 15-percentage-point drop in $\mathcal{D}_{\text{Held}}$.

These results reinforce the central claim of §3: the stability–plasticity tradeoff in factual fine-tuning is governed by how much new factual content the model integrates, not by the structural form of the constraint. Freezing is one mechanism for controlling that integration; self-distillation (§4.1) is another, achieving the same stability without sacrificing plasticity.

D Results Over Different Parameters for Self-distillation

Self-distillation introduces three hyperparameters: the snapshot epoch i at which the teacher is frozen, the distillation weight λ , and the temperature τ . Figure 10 reports ablations over each parameter independently: the snapshot-epoch sweep is run at $\lambda = 1$, $\tau = 2$, and the λ and τ sweeps at snapshot epoch $i = 2$ (with $\tau = 2$ and $\lambda = 1$, respectively). Each row shows accuracy on $\mathcal{D}_{\text{Unk}}$ (left, factual plasticity) and $\mathcal{D}_{\text{Held}}$ (right, factual stability) over training epochs.

Snapshot epoch. When the teacher is frozen at $i = 0$, $\mathcal{D}_{\text{Unk}}$ acquisition is substantially slower than under all other snapshot choices, and $\mathcal{D}_{\text{Held}}$ is the lowest across the ablation. In this setting, the distillation loss constrains the student toward a parameter space that predates task adaptation; since the model has yet to internalize the QA format, the lower $\mathcal{D}_{\text{Held}}$ may reflect difficulty in task learning rather than factual forgetting per se. Snapshot $i = 1$ yields the highest $\mathcal{D}_{\text{Held}}$ values, with the model already adapted to the task format but not yet subject



to sustained factual drift. Beyond $i = 1$, held-out stability degrades monotonically with the snapshot index.

Distillation weight. The λ ablation reveals a direct stability-plasticity tradeoff. Weaker regularization ($\lambda = 0.1, \lambda = 0.5$) allows rapid acquisition of $\mathcal{D}_{\text{Unk}}$ but results in greater degradation of $\mathcal{D}_{\text{Held}}$. Stronger regularization ($\lambda = 1.5$) substantially reduces forgetting but delays factual learning, requiring considerably more training epochs before $\mathcal{D}_{\text{Unk}}$ accuracy reaches its asymptote. We select $\lambda = 1$ as it achieves efficient factual acquisition alongside meaningful forgetting mitigation, without demanding an extended training budget.

Temperature. Higher temperatures ($\tau = 2, \tau = 4$) produce smoother and more stable learning trajectories on $\mathcal{D}_{\text{Unk}}$, reducing epoch-to-epoch variance. However, $\mathcal{D}_{\text{Held}}$ is noticeably lower under high- τ conditions. Lower temperatures ($\tau = 0.5, \tau = 1$) yield better stability on $\mathcal{D}_{\text{Held}}$ while preserving comparable final accuracy on $\mathcal{D}_{\text{Unk}}$. We adopt $\tau = 0.5$ as our default; practitioners who prioritize stable factual learning trajectories may prefer higher temperatures at a modest additional forgetting cost. The best-performing point of the sweep itself is ($i = 2, \lambda = 1, \tau = 1$). The default configuration used throughout the paper ($i = 1, \lambda = 1, \tau = 0.5$) retains $\mathcal{D}_{\text{Held}} = 0.908$ at epoch 50, about 2 pp below the 0.93 peak; this is the value reported for self-distillation in Table 5 and the source of the figure quoted in §4.1.

E Parameter-Space Baselines: Protocols and Sweeps

All baselines run in the 8K ENTITYQUESTIONS setting of §2 (50/50 Known/Unknown mixture, learning rate 5×10^{-5} , identical supervision and evaluation), trained to convergence. Table 5 summarizes each method at its best-performing hyperparameter.

Method	Final $\mathcal{D}_{\text{Held}}$	Final $\mathcal{D}_{\text{Unk}}$
Standard SFT	0.780	0.946
Weight decay (best strength: 50)	0.804	0.909
ℓ_2 toward θ_i	0.809	0.949
EWC ($\mathcal{D}_{\text{Known}}$ Fisher proxy, best λ)	0.826	0.948
LoRA (best rank: 64)	0.834	0.948
Self-distillation (ours)	0.908	0.934

Table 5: **Output-space vs. parameter-space regularization** in the 8K EntityQuestions setting (peak $\mathcal{D}_{\text{Held}} \approx 0.93$). Each baseline at its best-performing hyperparameter.

EWC. Standard EWC (Kirkpatrick et al., 2017) requires the old-task data (for the Fisher information), and SI (Zenke et al., 2017) requires the training trajectory (for online importance accumulation); neither is accessible for pretraining, which rules out SI entirely and requires a proxy for EWC. We therefore estimate the empirical Fisher on a sample of $\mathcal{D}_{\text{Known}}$ as a stand-in for the pretrained factual knowledge. Across regularization strengths, EWC reproduces the ℓ_2 pattern: the best-balanced setting reaches final $\mathcal{D}_{\text{Held}} = 0.826$ at full acquisition ($\mathcal{D}_{\text{Unk}} = 0.948$), while stronger regularization ($\lambda = 10^5$) raises $\mathcal{D}_{\text{Held}}$ to 0.869 only by collapsing acquisition ($\mathcal{D}_{\text{Unk}} = 0.444$).

LoRA. Low-rank adaptation (Hu et al., 2022) across ranks $\{16, 64\}$ applied to all modules. Consistent with Biderman et al. (2024), LoRA forgets somewhat less than full fine-tuning on a slower trajectory, but the pattern matches full SFT: final $\mathcal{D}_{\text{Held}} = 0.814$ (rank 16) and 0.834 (rank 64) at $\mathcal{D}_{\text{Unk}} \approx 0.95$.

Weight decay. AdamW decoupled weight decay across strengths $\{0.01, 0.1, 1, 10, 50, 100\}$. Moderate strengths leave forgetting essentially unchanged (final $\mathcal{D}_{\text{Held}} = 0.78\text{--}0.82$); the strongest setting (100) destroys training entirely. The best-performing strength (50) reaches $\mathcal{D}_{\text{Held}} = 0.804$ at $\mathcal{D}_{\text{Unk}} = 0.909$.

ℓ_2 toward θ_i . The snapshot-anchored ℓ_2 penalty of §5.2 with λ matched to self-distillation in gradient magnitude: final $\mathcal{D}_{\text{Held}} = 0.809$ at $\mathcal{D}_{\text{Unk}} = 0.949$.

Summary. Every parameter-space method lands in the 0.78–0.84 range on final $\mathcal{D}_{\text{Held}}$ (Table 5), a gain of at most ~ 5 pp over standard SFT versus ~ 13 pp for self-distillation, and each faces the same stability–plasticity tradeoff when its regularization strength increases. The self-distillation row uses the paper’s default configuration ($i = 1$, $\lambda = 1$, $\tau = 0.5$; Appendix D), evaluated at epoch 50. This is the pattern predicted by the associative-memory account (§5.3): weight-space penalties bound the update magnitude uniformly, without regard to the representational overlap ρ_{jA} that determines which stored facts are at risk.

F Real-Data Validation on Alpaca: Protocol and Trajectories

Data and annotation. We use 15K examples from the Alpaca dataset (Taori et al., 2023). Because factual content in instruction-tuning data is implicit (woven into instructions and responses rather than stated as clean relational facts), we first pass each sample through an LLM annotator (LLAMA-3.3-70B-Instruct via TogetherAI) that decides whether the sample carries factual knowledge and, if so, extracts the underlying (subject, relation, object) content. For each factual sample we generate an ENTITYQUESTIONS-style QA probe, and apply SLICK (§2.1) with QWEN 2.5-1.5B to classify the model’s knowledge state for that fact. A set of 1,000 factual probes is held out for evaluation; the held-out factual accuracy reported in Table 2 is computed over the *HighlyKnown* subset of these probes. We emphasize that no output of this pipeline is used during training: the mitigation methods receive only the raw Alpaca data, and SLICK serves purely as a diagnostic.

Training and evaluation. All three configurations fine-tune QWEN 2.5-1.5B for five epochs with learning rate 3×10^{-5} and batch size 16. Self-distillation uses the same hyperparameters as in §4.1 with no additional tuning; attention-only freezing follows §3. After every epoch we evaluate (i) held-out factual accuracy and (ii) alignment quality via AlpacaEval (Li et al., 2023) on 200 instructions, judged by LLAMA-3.3-70B-Instruct via TogetherAI against the dataset’s reference outputs; the preferred rate counts wins plus half of ties. Table 2 reports AlpacaEval at a common early-stopped epoch (4), the epoch selected by validation preferred rate for the two fully-trained configurations (standard SFT and self-distillation); we apply the same epoch to attention-only freezing for comparability, as its trajectory is essentially flat (its own best preferred rate, 50.3%, occurs at epoch 2). Factual degradation is reported at the end of training, measured against the peak held-out accuracy observed across all three runs (0.869, attained under standard SFT at epoch 3): $0.869 \rightarrow 0.795$ (SFT), 0.852 (self-distillation), and 0.860 (attention-only), i.e. 8, 2, and 1 pp at the two-decimal resolution appropriate for the 229-probe evaluation set (one probe ≈ 0.4 pp). AlpacaEval on 200 instructions has run-to-run variability of roughly $\pm 1\text{--}2$ pp: the same base model scored 39.8%, 38.8%, and 37.8% across the three runs’ initial evaluations, so differences of this order should be read with care.

Trajectories. Held-out factual accuracy over epochs 0–5 (epoch 0 = base model): standard SFT $0.821 \rightarrow 0.821 \rightarrow 0.843 \rightarrow 0.869 \rightarrow 0.803 \rightarrow 0.795$; self-distillation $0.821 \rightarrow 0.817 \rightarrow 0.834 \rightarrow 0.847 \rightarrow 0.843 \rightarrow 0.852$; attention-only freezing $0.817 \rightarrow 0.856 \rightarrow 0.852 \rightarrow 0.852 \rightarrow 0.847 \rightarrow 0.860$. The pattern matches the controlled setting: standard SFT peaks early and then degrades as factual content is absorbed, while both mitigations remain stable throughout. AlpacaEval preferred rates at epoch 4 are 50.5% (SFT), 52.3% (self-distillation), and 47.2% (attention-only), against a base-model rate of 39.8%.

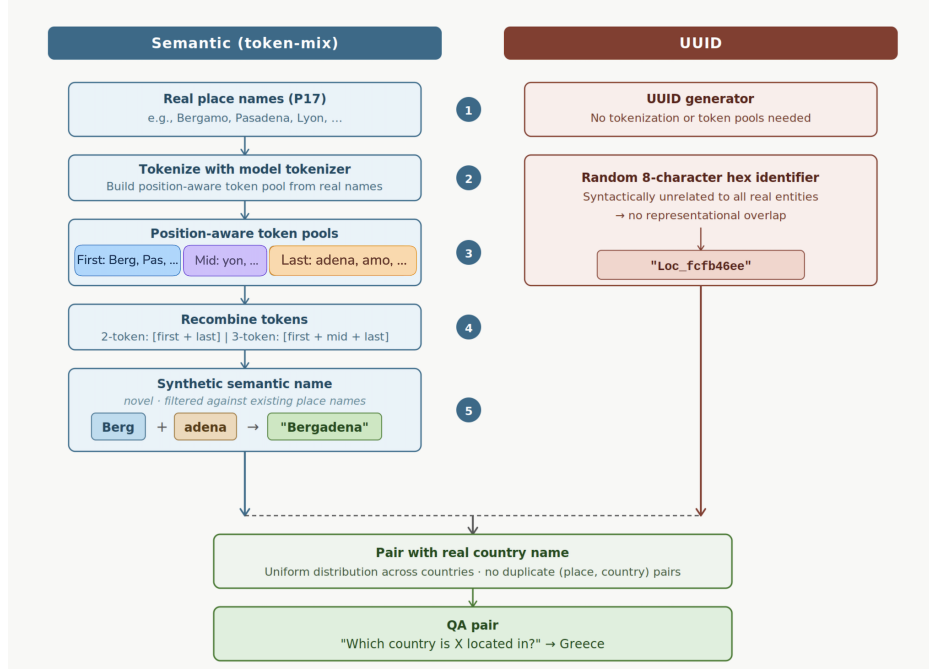


Figure 11: **Synthetic entity construction.** *Semantic (token-mix) keys* (left, steps 1–5): real P17 place names are tokenized, position-aware token pools are assembled, and novel names are formed by recombining tokens across names (e.g., **Bergadena** from **Bergamo** and **Pasadena**). *UUID keys* (right): random 8-character hex identifiers with no lexical overlap with real entities. Both conditions are paired with real country names under an identical question template, so the entity key is the sole variable. Candidate synthetic names that match any existing real place name are discarded.

G Synthetic Entities Creation Procedure

All controlled experiments in §5 use the P17 Wikidata relation (*location* → *country*: “Which country is X located in?”). We focus on a single relation deliberately: fixing the relation holds the question template, the answer domain, and the relational structure constant across all conditions, making the entity name (the *key*) the sole experimental variable and the comparison between key types experimentally controlled.

To construct synthetic facts, we generate two types of entity names (Fig. 11). **Semantic (token-mix) keys** are built by tokenizing all real P17 place names and extracting three position-aware token pools: the *first* token of each multi-token name, all *middle* tokens, and the *last* token. Novel place names are then formed by recombining tokens sampled from these pools, either as two-token sequences [first, last] or three-token sequences [first, middle, last], and any candidate that coincides with an existing real place is discarded. The resulting names share sub-word structure with real locations: for example, tokens from **Bergamo** and **Pasadena** combine into **Bergadena**, a novel name that lexically resembles genuine place names and is thus expected to activate overlapping representations in the model. **UUID keys**, by contrast, are random identifiers of the form `Loc_⟨8-hex⟩` (e.g., `Loc_fcfb46ee`), with no syntactic resemblance to any real entity. For the dose-response conditions of §5.1, we additionally generate **token-mix k -noise** keys ($k \in \{1, 2\}$) by replacing k of the tokens in each token-mix name with arbitrary identifier tokens drawn from the pool used for UUID construction, respecting token position; all other aspects of the construction (value assignment, filtering, balancing) are identical to the token-mix condition. In both conditions, each synthetic entity is paired with a real country drawn uniformly at random from those appearing in P17, subject to the constraint that the (place, country) pair is not already present in the original dataset; country assignments are rebalanced to maintain an approximately uniform distribution.

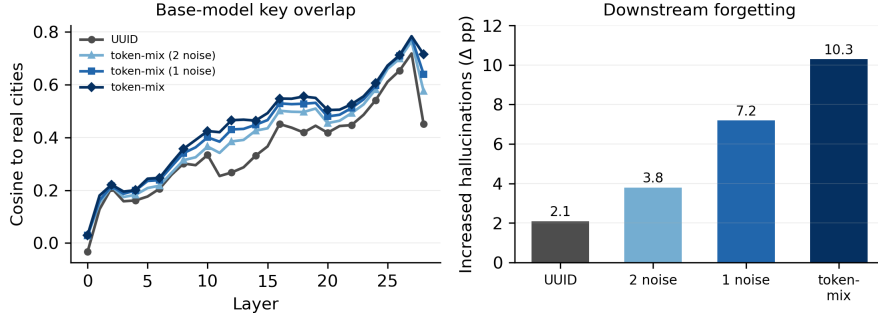


Figure 12: **Dose-response: forgetting tracks base-model representational overlap.** **Left:** per-layer cosine similarity of synthetic keys to real-city representations in the *base* model, four graded-overlap conditions. **Right:** forgetting at 10^4 injected facts increases monotonically with overlap.

To verify that forgetting is driven by the key rather than the value, we additionally varied the country (value) condition: values were instantiated as real country names, semantic synthetic country names (constructed by the same token-mix process applied to country names), or UUID-style country labels, crossed with both key types. Table 6 reports $\Delta \mathcal{D}_{\text{Held}}$ for all six combinations after training on 10^6 synthetic facts. Across all value conditions, semantic keys consistently induce 38–41 pp of forgetting, while UUID keys yield approximately 4 pp forgetting regardless of what value is paired with them. This full dissociation indicates that, among the factors varied here, representational overlap in the entity name is the dominant driver of interference, while the answer side plays little role.

Key type	Value type	$\Delta \mathcal{D}_{\text{Held}}$ (pp) ↓
Semantic	Real	38
Semantic	Semantic	38
Semantic	UUID	41
UUID	Real	4
UUID	Semantic	4
UUID	UUID	4

Table 6: **Forgetting is driven by key semantics, not value type.** $\Delta \mathcal{D}_{\text{Held}}$ after training on 10^6 synthetic facts, for all combinations of key type (semantic vs. UUID) and value type (real / semantic / UUID country names). All conditions achieve $\mathcal{D}_{\text{Unk}} = 100\%$ and $\mathcal{D}_{\text{Known}} \geq 90\%$, so differences in forgetting cannot be attributed to learning success.

H Representational Geometry of Synthetic Keys

Setup. For each key condition (real cities, token-mix, token-mix 1-noise, token-mix 2-noise, UUID) we extract residual-stream representations of $n = 5,000$ entities per condition from the *base* QWEN 2.5-1.5B model (no fine-tuning), at every layer, under two aggregations: the mean over the entity’s sub-word tokens and the last sub-word token. Results below use the last-sub-word aggregation, which matches how key representations are read in §5.2; the mean aggregation agrees at the early and final layers but saturates mid-network, where all conditions approach ceiling similarity. Two additional runs ($n = 2,000$, two seeds) replicate the cosine ordering at every layer; the probe ordering at this smaller scale is reproduced in one seed and flat in the other.

Cosine similarity to real cities. Figure 12 (left) reports the mean pairwise cosine similarity between entities of each synthetic condition and real cities. The ordering $\text{UUID} < 2\text{-noise} < 1\text{-noise} < \text{token-mix}$ holds at every layer past the earliest ones, with last-layer values 0.45, 0.58, 0.64, 0.71. Within-condition similarity to *real-real* pairs behaves analogously.

Linear “is a city” probe. We train an ℓ_2 -regularized logistic probe on base-model representations to separate real cities from non-place entities, and evaluate the predicted $P(\text{city})$ on each synthetic condition. At layer 27, the deepest layer at which the probe is reliable, it transfers with the same ordering as the cosine analysis: real 0.91, token-mix 0.71, 1-noise 0.62, 2-noise 0.50, UUID 0.29. The clean graded ordering holds in the final layers (25–28); at mid-network layers the probe itself is unreliable, with the UUID condition in particular oscillating widely. The cosine measure instead orders the conditions monotonically at every layer.

Takeaway. Both measures agree that the graded key conditions occupy graded positions in the base model’s representation space before any fine-tuning, so the dose-response in forgetting (Fig. 5) tracks pre-existing representational overlap rather than an artifact of training. Forgetting values follow the convention of Fig. 4, measuring the drop from the ≈ 0.98 plateau reached after task-format learning.

I Further Analyses on Self-Distillation

We report two sets of additional analyses that complement §5.2, both using the 10K semantic-overlap setting of §5.1: ablations on which part of the output distribution the distillation constraint needs to target, and an extended battery of drift metrics comparing SFT and self-distillation across representation and output space.

Which part of the output distribution matters? The main text establishes that ℓ_2 weight regularization does not replicate the self-distillation benefit, suggesting the effect is not simply a consequence of smaller weight updates. We additionally ask whether the *full* output distribution is necessary, or whether constraining a specific region suffices. This question connects to a broader literature on knowledge distillation: Hinton et al. (2015) showed that the “dark knowledge” carried by a teacher’s soft probability distribution over non-target classes is more informative than the hard labels alone, precisely because it encodes the teacher’s relative confidence over semantically related alternatives. In our setting, those alternatives are competing entity tokens — and we hypothesize that it is constraining *this* region, rather than the full distribution, that drives the forgetting reduction. We compare three variants of the distillation objective:

Full self-distillation: KL divergence over the complete vocabulary at each token position (Eq. (2)).

Top- k distillation: KL divergence restricted to the teacher’s k highest-probability tokens at each position. Let $T_k(j) = \arg \text{top}_k p_{\theta_i}(\cdot \mid x_{<j})$ be the top- k index set at position j , and let $p_{\theta_i,j}^{(k)}, p_{\theta,j}^{(k)}$ be the student and teacher distributions renormalized over $T_k(j)$. The distillation loss becomes

$$\mathcal{L}_{\text{distill}}^{(k)}(\theta; \theta_i) = \mathbb{E}_{(x,y) \sim \mathcal{B}} \left[\frac{\tau^2}{|M(y)|} \sum_{j \in M(y)} \text{KL}(p_{\theta_i,j}^{(k)} \parallel p_{\theta,j}^{(k)}) \right]. \quad (5)$$

We set $k = 0.05\%$ of the vocabulary (76 tokens), capturing between 91% and 92% of the teacher’s probability mass per position.

Random- k distillation: identical to top- k , but the 76 tokens at each position are drawn uniformly at random rather than by probability rank. This condition controls for the *number* of constrained logits while removing alignment with the teacher’s high-probability region. Importantly, it does not reduce to standard SFT: the KL term remains active but constrains the student toward a uniform distribution over an arbitrary token subset, effectively injecting noise rather than meaningful structure.

Top- k distillation fully replicates full self-distillation: forgetting on $\mathcal{D}_{\text{Held}}$ decreases to ≈ 3 pp while $\mathcal{D}_{\text{Unk}}$ acquisition proceeds at the same pace. Random- k distillation leaves forgetting at ≈ 10 pp, indistinguishable from standard SFT. What matters is therefore not the number of constrained logits but constraining the high-probability region specifically. This is consistent

with Hinton et al. (2015)’s insight that soft targets carry informative relational structure — here, the relative probabilities over competing entity tokens encode the teacher’s “opinion” about which alternatives are plausible, and preserving this opinion prevents the gradient updates for new entities from redistributing probability mass away from existing ones. Buzzega et al. (2020); Li et al. (2025) make a related observation in the continual learning setting, showing that dark experience replay — which stores and replays soft teacher outputs rather than hard labels — provides a stronger anti-forgetting signal than hard-label replay alone, precisely because the soft outputs encode richer relational structure. Our top- k result sharpens this: what matters is not the full soft distribution, but the high-probability region where semantically competing candidates reside.

Notation for drift metrics. Let $H_\theta^{(l)}(x) \in \mathbb{R}^d$ denote the hidden representation of the key entity’s final token at transformer layer l under model θ , for input x . We write $H_0^{(l)}(x) = H_{\theta_0}^{(l)}(x)$ for the pretrained model and $H_i^{(l)}(x)$ for epoch i . All representations are extracted from the final token of the key entity and ℓ_2 -normalized before computing cosine similarities. The final token is the natural extraction point for entity representations in causal language models: unlike bidirectional architectures, where the full contextual representation of a phrase can reside at any of its token positions (Kaplan et al., 2026), causal attention ensures that only the last token has attended to all preceding tokens and thus carries the complete representation of the entity. We fix $l = 14$: the middle layer of the 28-layer Qwen 2.5-1.5B model, where the inner lexicon of LLMs is most reliably encoded (Kaplan et al., 2025). Let $\mathcal{P} = \{(x_u, x_h) : x_u \in \mathcal{D}_{\text{Unk}}, x_h \in \mathcal{D}_{\text{Held}}, a(x_u) = a(x_h)\}$ be the set of unknown/held-out input pairs sharing the same ground-truth answer $a(\cdot)$, and $I = T_k^{\theta_0}(x_h)$ the pretrained model’s top- k token indices at the answer position for held-out input x_h .

The battery of metrics below is designed to triangulate the interference account from multiple angles: RD measures whether individual held-out representations move; NS measures whether their *relative geometry* is preserved; MTD measures whether they drift specifically *toward* new-entity representations; and Rank- ρ , NTR, and JSD measure whether this interference is visible at the output level. Convergent evidence across all six metrics strengthens the structural account over capacity-based or behavioral alternatives.

Hidden-state drift (RD). RD (defined in §5.2) measures the average cosine distance between held-out entity representations before and after fine-tuning. Cosine distance captures directional shifts in representation space, which are more sensitive to semantic change than Euclidean distance (Kim et al., 2025). Under SFT, RD reaches $\approx 11\%$; under self-distillation it stabilizes near 5%. The residual 5% present in all conditions reflects task-format learning and is not associated with forgetting, as confirmed by the UUID condition in §5.1.

Neighborhood structure (NS).

$$\text{NS}_i = \frac{1}{|\mathcal{S}|^2} \sum_{x, x' \in \mathcal{S}} |\cos(H_i^{(14)}(x), H_i^{(14)}(x')) - \cos(H_0^{(14)}(x), H_0^{(14)}(x'))|, \quad (6)$$

where $\mathcal{S} \subseteq \mathcal{D}_{\text{Held}}$ is a random subsample. NS measures how much the pairwise geometric structure among held-out entity representations changes during training — capturing whether entities move *together* (high RD, low NS) or *rearrange relative to each other* (high NS). Under SFT, NS reaches ≈ 0.07 ; under self-distillation it stabilizes near 0.02. The fact that both RD and NS are elevated under SFT indicates that held-out representations are not merely translating as a block, but genuinely reorganizing — consistent with localized interference rather than a global drift of all representations. This pattern mirrors findings in Masip et al. (2026), who show that catastrophic forgetting in classification networks manifests as reorganization concentrated among semantically similar classes, rather than uniform drift.

Mean targeted drift (MTD).

$$\text{MTD}_i = \mathbb{E}_{(x_h, x_s)} [\cos(H_i^{(14)}(x_h), H_i^{(14)}(x_s)) - \cos(H_0^{(14)}(x_h), H_0^{(14)}(x_s))], \quad (7)$$

where $x_h \in \mathcal{D}_{\text{Held}}$ and x_s indexes the synthetic (newly trained) entities. MTD directly tests the directionality of the interference: does the representational drift push held-out

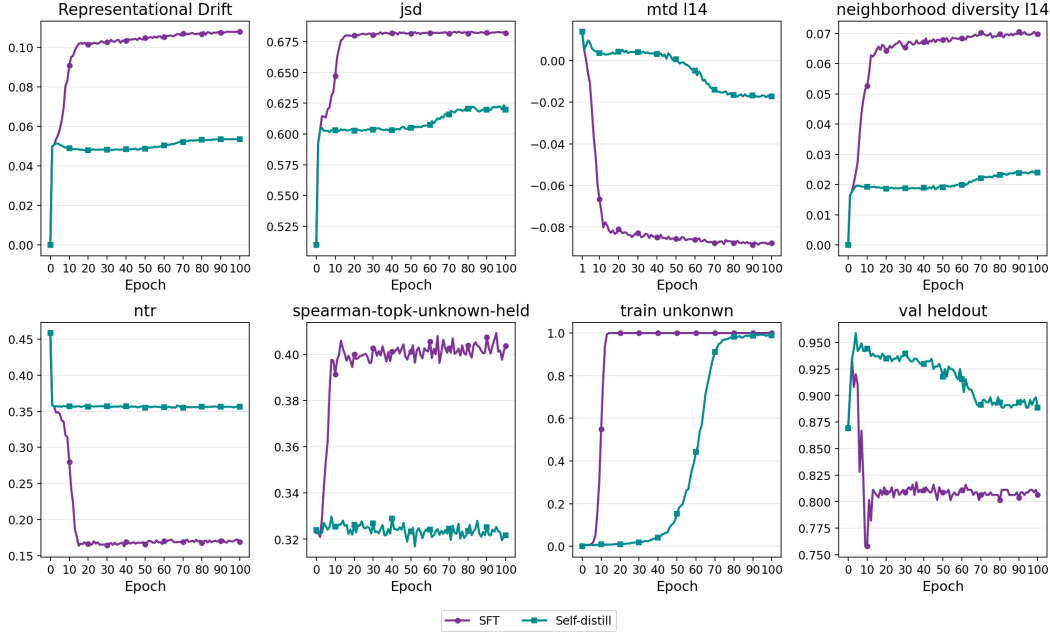


Figure 13: **Extended drift metrics under standard SFT and self-distillation** over 100 training epochs on the 10K semantic-overlap setting (§5.1). *Top row, left to right*: Hidden-state drift (RD), Jensen-Shannon divergence (JSD), mean targeted drift (MTD), and neighborhood structure (NS) — all at layer 14. *Bottom row, left to right*: Neighborhood token ratio (NTR), output-distribution drift (Rank- ρ), training accuracy on $\mathcal{D}_{\text{Unk}}$, and validation accuracy on $\mathcal{D}_{\text{Held}}$. All metrics consistently show less drift under **self-distillation** than under **standard SFT**. Factual acquisition on $\mathcal{D}_{\text{Unk}}$ proceeds at a comparable pace across conditions.

entities *toward* or *away from* the new entities? Under SFT, MTD drops to ≈ -0.09 : held-out and synthetic representations diverge — consistent with the model carving out distinct new-entity clusters that displace nearby held-out representations sideways rather than absorbing them. Under self-distillation, MTD remains near 0, indicating that the cross-group representational relationship is largely preserved. This divergence pattern is consistent with Nishi et al. (2025), who show that knowledge editing causes “representation shattering” — a reorganization of the local neighborhood structure around edited entities — and that nearby unedited entities are displaced as a byproduct.

Output-distribution drift (Rank- ρ). For each pair $(x_u, x_h) \in \mathcal{P}$, Rank- ρ measures the Spearman rank correlation between the current and pretrained models’ logit rankings over unknown-fact and held-out inputs, restricted to the shared index set I :

$$\text{Rank-}\rho = \frac{1}{|\mathcal{P}|} \sum_{(x_u, x_h) \in \mathcal{P}} \rho_{\text{Spearman}}(z_\theta(x_u)_I, z_{\theta_0}(x_h)_I), \quad (8)$$

where $z_\theta(x)_I$ denotes the logits of θ at the answer position restricted to I . A rising Rank- ρ indicates that the model’s output for new-fact inputs increasingly ranks entity candidates in the same order as the pretrained model ranked them for held-out inputs — a signature of the two entity neighborhoods converging in output space, which is precisely the interference pattern the structural account predicts. Under SFT, Rank- ρ rises from ≈ 0.32 to ≈ 0.40 ; under self-distillation it stays near 0.32.

Neighborhood token ratio (NTR).

$$\text{NTR} = \mathbb{E}_{(x_u, x_h) \in \mathcal{P}} \left[\frac{|T_k^\theta(x_u) \cap T_k^{\theta_0}(x_h)|}{k} \right], \quad (9)$$

where $T_k^\theta(x)$ is the current model’s top- k token set at the answer position for input x . NTR complements Rank- ρ by measuring overlap in the *current* model’s candidate sets (rather

than correlation with the pretrained reference): do new-fact queries and held-out queries share the same answer candidates under the fine-tuned model? At initialization, the two groups share $\approx 45\%$ of their top- k candidates. Under SFT, NTR drops to $\approx 17\%$ as the model learns specialized answer distributions for the new entities; under self-distillation it remains near 36%. The drop under SFT reflects the model pulling new-entity outputs into a distinct region of output space — consistent with the MTD finding that held-out and synthetic representations diverge.

Jensen-Shannon divergence (JSD).

$$\text{JSD} = \mathbb{E}_{(x_u, x_h) \in \mathcal{P}} [\text{JSD}(p_\theta(x_u)_I \parallel p_{\theta_0}(x_h)_I)], \quad (10)$$

where $p_\theta(x)_I$ is the softmax distribution of θ at the answer position renormalized over I . JSD captures differences in the full probability shape rather than just rankings (complementing Rank- ρ). Both conditions show an upward trend, as the model naturally diverges from the fixed pretrained reference as it learns new facts. However, SFT reaches ≈ 0.68 versus ≈ 0.63 under self-distillation. The smaller gap here compared to Rank- ρ and NTR reflects the fact that JSD is sensitive to all probability mass, including the task-format learning component shared across conditions; Rank- ρ and NTR, by focusing on the entity-candidate region, isolate the interference-specific signal more cleanly.

Training and evaluation curves. The *train unknown* and *val heldout* subplots in Fig. 13 confirm that both conditions achieve comparable factual acquisition on $\mathcal{D}_{\text{Unk}}$ — self-distillation converges somewhat later — while held-out accuracy on $\mathcal{D}_{\text{Held}}$ is substantially better preserved under self-distillation, consistent with the main results of §4.1. Taken together, all six metrics tell a consistent story: SFT on semantically overlapping entities produces representational and output-space interference that is absent under self-distillation, consistent with this interference being a primary driver of factual forgetting.

J Derivations for the Hebbian Associative-Memory Account

This appendix collects the derivations referenced in §5.3. Each subsection is self-contained and can be read independently.

Notation. The main text needs only the held-out subject’s key, and writes x_h for that subject. The derivations below also involve its stored answer and its margin, so we label the held-out *fact* by A , with subject x_A , answer a_A , and key $k_A := k_{x_A}$; the correspondence is $x_A = x_h$, $k_A = k_{x_h}$, $\rho_{jA} = \rho_{jh}$, and $S_A = S_h$, and $\mathcal{A}$ denotes the answer set over which the free index y ranges. Eq. (4) gives the effect of a single injection j ; $\delta_y(A) = \sum_{j=1}^M \delta_y^{(j)}(x_h)$ below is the cumulative perturbation from all M injections, which §5.3 describes in prose.

J.1 Hebbian update derivation

Lemma 1 (SFT update is a rank-one Hebbian increment). *Consider one gradient step of softmax cross-entropy on a new fact (x_j, a_j) with learning rate η , treating W in Eq. (3) as a free linear readout and holding the key map $x \mapsto k_x$ fixed. The induced update is*

$$\Delta W_j = \eta (u_{a_j} - \bar{u}_j) k_{x_j}^\top, \quad \bar{u}_j := \sum_{y \in \mathcal{A}} p_y(x_j) u_y, \quad p(x_j) = \text{softmax}(s(x_j)). \quad (11)$$

Proof. Let $\mathcal{L}_j$ denote the softmax-cross-entropy loss on (x_j, a_j) with logits $s_y = u_y^\top W k_{x_j}$. The gradient with respect to logits is

$$\frac{\partial \mathcal{L}_j}{\partial s_y} = p_y(x_j) - \mathbf{1}[y = a_j].$$

By the chain rule,

$$\nabla_W \mathcal{L}_j = \sum_{y \in \mathcal{A}} (p_y(x_j) - \mathbf{1}[y = a_j]) u_y k_{x_j}^\top = (\bar{u}_j - u_{a_j}) k_{x_j}^\top,$$

and $\Delta W_j = -\eta \nabla_W \mathcal{L}_j$ gives (11). $\square$

The identification is structural: the update (11) is exactly a rank-one outer-product increment, the same primitive that builds the stored memory $W = \sum_x u_{f^*(x)} k_x^\top$ in Eq. (3). SFT on M new facts yields, to first order, $W' = W + \sum_{j=1}^M \Delta W_j$. The two approximations are (i) first order in η (per-example rank-ones rather than a fully entangled trajectory), and (ii) the key map is held fixed during fine-tuning.

J.2 Bias and variance of $\delta_y(A)$

We work under three conditions on the SFT corpus, the first two structural and the third a mild regularity:

- (a) (Disjoint held-out answer.) $a_A \notin \{a_j\}_{j=1}^M$.
- (b) (Distinct injected answers.) $a_j \neq a_{j'}$ for $j \neq j'$.
- (R) (Cross-prior regularity.) There exists $C_R \geq 0$ such that for every $j \in [M]$ and every $a \in \mathcal{A}$,

$$\sum_{j' \neq j} p_a(x_{j'}) \leq C_R. \quad (12)$$

Conditions (a) and (b) are idealizations of our experimental corpora: in the synthetic settings, values are country names that repeat across injected facts, and held-out answers are drawn from the same set, so both conditions are violated in practice. We adopt them to isolate the overlap mechanism in its cleanest form. With repeated answers, the Gram matrix B below acquires additional unit off-diagonal entries whose row sums grow with the per-answer repetition count; the bound then holds with a larger constant that depends on this count, and the S_A -monotone form of the variance is unchanged at any fixed repetition profile. Condition (R) says that the total prior mass that any single answer accumulates across the other $M-1$ new subjects is bounded; a sufficient condition is $\max_{j,a} p_a(x_j) \leq C_R/M$, which is plausible in our regime since injected facts are *Unknown* (the model assigns low probability even to their target answers before fine-tuning).

Proposition 1 (Bias-variance decomposition). *Define $\delta_y(A) := u_y^\top (W' - W) k_A$, the perturbation of the readout score of answer y on the held-out subject x_A , and write $\delta_y(A) = \mu_y(A) + \xi_y(A)$ with $\mu_y(A) := \mathbb{E}_u[\delta_y(A)]$ and $\xi_y(A) := \delta_y(A) - \mu_y(A)$, where the expectation is over random-spherical unembeddings (u_a unit-norm with $\mathbb{E}[u_a u_a^\top] = I_d/d$).*

(a) Bias (exact). Under (a) and (b), for every $y \in \mathcal{A}$,

$$\mu_y(A) = \eta \sum_{j=1}^M (\mathbf{1}[y = a_j] - p_y(x_j)) \rho_{jA}. \quad (13)$$

In particular, for $y = a_j$ (an injected answer),

$$\mu_{a_j}(A) = \eta (1 - p_{a_j}(x_j)) \rho_{jA} - \eta \sum_{j' \neq j} p_{a_j}(x_{j'}) \rho_{j'A}, \quad (14)$$

and for $y \notin \{a_j\}_{j=1}^M$,

$$\mu_y(A) = -\eta \sum_{j=1}^M p_y(x_j) \rho_{jA}. \quad (15)$$

The diagonal term in (14) is the systematic “pull” of fact j on the readout for A , linear in ρ_{jA} , with prefactor $1 - p_{a_j}(x_j) \in (0, 1)$ that vanishes only as SFT saturates $p_{a_j}(x_j) \rightarrow 1$. Under (R), the cross term in (14) and the bias (15) are bounded in absolute value by $\eta (1 + C_R) \max_j |\rho_{jA}|$.

(b) Variance (upper bound). Under (a), (b), and (R), for every $y \in \mathcal{A}$,

$$\text{Var}_u[\xi_y(A)] \leq \frac{C \eta^2}{d} S_A, \quad S_A := \sum_{j=1}^M \rho_{jA}^2 = M \overline{\rho_A^2}, \quad C \leq 2C_R + 3. \quad (16)$$

Proof. Regrouping by answer token. Reindex the cumulative update $W' - W = \sum_{j=1}^M \Delta W_j$ (with ΔW_j from Eq. (11)) by answer rather than by fact:

$$W' - W = \eta \sum_{a \in \mathcal{A}} u_a v_a^\top, \quad v_a := \sum_{j=1}^M \alpha_{a,j} k_{x_j}, \quad \alpha_{a,j} := \mathbf{1}[a = a_j] - p_a(x_j). \quad (17)$$

Substituting,

$$\delta_y(A) = \eta \sum_{a \in \mathcal{A}} (u_y^\top u_a) (v_a^\top k_A). \quad (18)$$

Bias. Under the unembedding isotropy, $\mathbb{E}_u[u_y^\top u_a] = \mathbf{1}[y = a]$. Taking expectations of (18),

$$\mu_y(A) = \eta v_y^\top k_A = \eta \sum_{j=1}^M \alpha_{y,j} \rho_{jA},$$

which is (13). The special cases (14) and (15) follow by substituting $y = a_j$ (condition (b) collapses $\mathbf{1}[a_j = a_{j'}]$ to $j' = j$) and $y \notin \{a_j\}$ (the indicator vanishes everywhere). For the bounds stated below the proposition: by the L^1/L^∞ inequality and condition (R),

$$\left| \sum_j p_y(x_j) \rho_{jA} \right| \leq \left(\sum_j p_y(x_j) \right) \max_j |\rho_{jA}| \leq (1 + C_R) \max_j |\rho_{jA}|,$$

using $\sum_j p_y(x_j) = p_y(x_{j_0}) + \sum_{j \neq j_0} p_y(x_j) \leq 1 + C_R$ for any j_0 .

Variance: reduction to a Gram matrix. Subtracting the mean from (18),

$$\xi_y(A) = \eta \sum_{a \neq y} (u_y^\top u_a) (v_a^\top k_A)$$

(the $a = y$ term is exactly μ_y). Under isotropy, $\{u_y^\top u_a\}_{a \neq y}$ are pairwise uncorrelated with $\text{Var}_u[u_y^\top u_a] = 1/d$, hence

$$\text{Var}_u[\xi_y(A)] = \frac{\eta^2}{d} \sum_{a \neq y} (v_a^\top k_A)^2 \leq \frac{\eta^2}{d} \|V k_A\|^2, \quad (19)$$

where V is the operator with columns $\{v_a\}_{a \in \mathcal{A}}$. Expanding the squared norm,

$$\|V k_A\|^2 = \sum_{a \in \mathcal{A}} (v_a^\top k_A)^2 = \sum_{j,j'=1}^M \rho_{jA} \rho_{j'A} B_{jj'} = \rho_A^\top B \rho_A, \quad (20)$$

where $\rho_A := (\rho_{1A}, \dots, \rho_{MA})^\top$ and $B \in \mathbb{R}^{M \times M}$ is the Gram matrix

$$B_{jj'} := \sum_{a \in \mathcal{A}} \alpha_{a,j} \alpha_{a,j'} = \mathbf{1}[a_j = a_{j'}] - p_{a_j}(x_{j'}) - p_{a_{j'}}(x_j) + \langle p(x_j), p(x_{j'}) \rangle. \quad (21)$$

The last equality expands $\alpha_{a,j} \alpha_{a,j'}$ and sums over a (the indicator product collapses to $\mathbf{1}[a_j = a_{j'}]$ because both indicators are simultaneously 1 only when $a = a_j = a_{j'}$). Since B is symmetric PSD (a Gram matrix), $\rho_A^\top B \rho_A \leq \|B\|_{\text{op}} \|\rho_A\|_2^2 = \|B\|_{\text{op}} S_A$.

Bounding $\|B\|_{\text{op}}$ via Gershgorin. On the diagonal, $\mathbf{1}[a_j = a_j] = 1$ and $\|p(x_j)\|_2^2 \leq \max_a p_a(x_j) \cdot \sum_a p_a(x_j) \leq 1$, so

$$B_{jj} = 1 - 2p_{a_j}(x_j) + \|p(x_j)\|_2^2 \leq 2.$$

Off the diagonal, condition (b) gives $\mathbf{1}[a_j = a_{j'}] = 0$ for $j \neq j'$, and the three remaining terms are nonnegative (probabilities and inner products of probability vectors), so

$$|B_{jj'}| \leq p_{a_j}(x_{j'}) + p_{a_{j'}}(x_j) + \langle p(x_j), p(x_{j'}) \rangle.$$

Summing $|B_{jj'}|$ over $j' \neq j$ and applying (R) three times:

$$\begin{aligned} \sum_{j' \neq j} p_{a_j}(x_{j'}) &\leq C_R \quad (\text{by (R) with } a = a_j), \\ \sum_{j' \neq j} p_{a_{j'}}(x_j) &\leq \sum_{a \in \mathcal{A}} p_a(x_j) = 1 \quad (\text{by (b), the } a_{j'} \text{ are distinct in } \mathcal{A}), \\ \sum_{j' \neq j} \langle p(x_j), p(x_{j'}) \rangle &= \sum_{a \in \mathcal{A}} p_a(x_j) \sum_{j' \neq j} p_a(x_{j'}) \leq C_R \sum_a p_a(x_j) = C_R \quad (\text{by (R)}). \end{aligned}$$

Hence $\sum_{j' \neq j} |B_{jj'}| \leq 2C_R + 1$. Since B is symmetric PSD, Gershgorin's circle theorem gives

$$\|B\|_{\text{op}} = \lambda_{\max}(B) \leq \max_j \left(B_{jj} + \sum_{j' \neq j} |B_{jj'}| \right) \leq 2 + (2C_R + 1) = 2C_R + 3.$$

Combining (19) and (20),

$$\text{Var}_u[\xi_y(A)] \leq \frac{\eta^2}{d} \|B\|_{\text{op}} S_A \leq \frac{(2C_R + 3) \eta^2}{d} S_A,$$

which is (16) with $C \leq 2C_R + 3$. $\square$

J.3 Forgetting probability and scaling law (heuristic)

Let $m_A := s_{a_A}(x_A) - \max_{y \neq a_A} s_y(x_A) > 0$ be the pre-finetuning margin of fact A . Fact A is forgotten when, for some competing answer $y \neq a_A$,

$$\delta_y(A) - \delta_{a_A}(A) > m_A. \quad (22)$$

By Proposition 1(a), $\mu_y(A)$ is nonzero (to leading order in (R)) only for $y \in \{a_j\}_{j=1}^M$; the maximum of $\mu_y(A)$ over $y \neq a_A$ is therefore attained at some $y = a_j$. Writing $\delta = \mu + \xi$, the forgetting condition (22) specializes to

$$\mu_{a_j}(A) - \mu_{a_A}(A) + (\xi_{a_j}(A) - \xi_{a_A}(A)) > m_A. \quad (23)$$

Variance of the difference (rigorous). By Proposition 1(b) applied to $y = a_j$ and $y = a_A$ separately, and the elementary inequality $\text{Var}[X - Y] \leq 2\text{Var}[X] + 2\text{Var}[Y]$,

$$\text{Var}_u[\xi_{a_j}(A) - \xi_{a_A}(A)] \leq \frac{4(2C_R + 3) \eta^2}{d} S_A. \quad (24)$$

Gaussian heuristic. By the proof of Proposition 1, $\xi_y(A) = \eta \sum_{a \neq y} (u_y^\top u_a)(v_a^\top k_A)$ is a linear combination of $|\mathcal{A}| - 1$ pairwise-uncorrelated mean-zero terms (the cross-products $u_y^\top u_a$ under isotropy), each of order $1/\sqrt{d}$. For large $|\mathcal{A}|$, a CLT approximation gives that $\xi_{a_j}(A) - \xi_{a_A}(A)$ is approximately Gaussian. Combining with (23) and (24), the forgetting probability is then a Gaussian tail:

$$\text{Pr}[\text{forget } A] \approx g\left(\frac{m_A - \max_j (\mu_{a_j}(A) - \mu_{a_A}(A))}{\eta \sqrt{4(2C_R + 3) S_A / d}}\right), \quad (25)$$

with g a decreasing function. This is the one heuristic step in the analysis: the Gaussian approximation is what justifies a specific functional form. We use (25) only to derive monotonic orderings (both the numerator decreases and the denominator increases as S_A grows), never to read off a specific exponent. The compact statement in §5.3 suppresses the bias term in the numerator (valid in the small- μ regime) and absorbs the constants into g ; the full form (25) is monotone in S_A either way.

J.4 Why output-space regularization beats weight-space regularization

The interference identified in §5.3 lives along the unembedding directions: by Proposition 1, both the bias $\mu_y(A)$ and the fluctuation $\xi_y(A)$ are perturbations of output logits. A KL penalty toward the pre-fine-tuning model therefore suppresses both directly, including on held-out subjects, which share the training subjects' representational geometry. A weight-space penalty instead bounds $\|\Delta W_j\|_F = \eta \|u_{a_j} - \bar{u}_j\| \|k_{x_j}\|$, a quantity independent of the held-out key k_A and hence blind to the overlap ρ_{jA} : it protects all stored facts uniformly rather than preferentially protecting the overlap-vulnerable ones, consistent with the ℓ_2 comparison of §5.2 and the parameter-space baselines of Table 5. The same logic explains why top- k distillation succeeds while random- k does not (Appendix I): by Eq. (13), the interference bias is nonzero, to leading order under condition (R), only for the injected answers $y \in \{a_j\}_{j=1}^M$, exactly the high-probability directions that top- k constrains, while random- k spends most of its budget on directions carrying no bias to cancel.

J.5 Extension to distributed storage

The main text models factual storage as a single effective associative memory W acting on the subject's internal representation k_x (Eq. (3)) without committing to a specific layer at which k_x is read. In real transformers, factual storage is distributed across multiple attention and MLP components, each potentially reading the residual stream at a different layer (Meng et al., 2022; 2023). This appendix shows that the main-text analysis extends to this distributed setting with no qualitative change.

Let component c read at layer ℓ_c with key $k_x^{(c)} := \phi_{\ell_c}(x)$, and let the readout decompose as a weighted sum $s_y(x) = \sum_c w_c u_y^\top W^{(c)} k_x^{(c)}$ for nonnegative component weights $w_c \geq 0$.

Lemma 1 applies component-wise: $\Delta W_j^{(c)} = \eta(u_{a_j} - \bar{u}_j)(k_{x_j}^{(c)})^\top$, so the perturbation is

$$\delta_y(A) = \sum_c w_c \eta \sum_{j=1}^M (u_y^\top (u_{a_j} - \bar{u}_j)) \rho_{jA}^{(c)}, \quad \rho_{jA}^{(c)} := \langle k_{x_j}^{(c)}, k_A^{(c)} \rangle. \quad (26)$$

The bias-variance decomposition of Proposition 1 applies term by term. Treating the components as independent in u (a simplification; cross-component covariances share the same u_a 's and are bounded by Cauchy-Schwarz), the variance bound becomes

$$\text{Var}_u[\xi_y(A)] \leq \frac{C\eta^2}{d} \sum_c w_c^2 \sum_{j=1}^M (\rho_{jA}^{(c)})^2. \quad (27)$$

Without the independence simplification the worst-case bound takes the form $(C\eta^2/d)(\sum_c w_c \sqrt{S_A^{(c)}})^2$, which is still monotone in each component overlap $S_A^{(c)} := \sum_j (\rho_{jA}^{(c)})^2$. The functional form of monotonicity in summed squared overlap is preserved either way, with overlaps now measured at the layer of each component. We do not pin down the weights w_c or the per-component read layers, which depend on which component dominates retrieval; what is preserved is the qualitative law.