Knowing but Not Saying: Preventing Factual Access Failures in LLM SFT via Recall-Anchored Distillation

Haodong Chen[1*], **Yadong Wang**[1*], **Shengtao Wen**[1], **Dong Liang**[1], **Xiang Chen**[1†]

[1]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence,
College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics
{haodong_chen,xiang_chen}@nuaa.edu.cn

Abstract

Supervised fine-tuning (SFT) can degrade factual behavior outside the target domain. This degradation is often described as catastrophic forgetting, yet open-ended factual failures do not necessarily imply that the underlying facts have been erased. In this work, we identify a more specific phenomenon, factual access failure: after domain SFT, models can still recognize or rank the correct answer under constrained evaluation, while failing to produce it in closed-book generation. Through benchmark-level comparisons, same-fact multiple-choice and generation probes, and failure-mode analysis, we show that SFT-induced factual degradation reflects both genuine wrong-answer generations and expression-level failures such as verbosity, formatting mismatch, and exact-match artifacts. To address this problem, we introduce Recall-Anchored Distillation (RAD), a base-anchored self-distillation objective that preserves out-of-distribution generation behavior by aligning the adapted model with the original base model's soft continuation distribution on unlabeled OOD text. RAD requires no gold OOD answers, external judges, or labeled factual data. Across three backbones fine-tuned on MedMCQA, RAD recovers a consistent portion of the lost OOD recall while preserving target-domain adaptation. Compared with replay on the same OOD text, RAD shows that the key preservation signal is the base model's soft distribution rather than additional text exposure alone.

Introduction

Supervised fine-tuning (SFT) is a standard approach for adapting large language models (LLMs) to specific tasks, domains, and output formats, including instruction following, domain-specific question answering, and structured response generation (Ouyang et al. 2022; Chung et al. 2022; Hu et al. 2021; Zhang et al. 2025; Yang et al. 2024a). In high-stakes domains such as medicine, this adaptation is often necessary because a base model must learn the format, terminology, and decision boundaries of the target task (Pal, Umapathi, and Sankarasubbu 2022; Singhal et al. 2023; Savage et al. 2024; Li, Wang, and Yu 2024). Yet SFT is not neutral: it can alter the topic preferences, stylistic behavior, factual behavior, and use of pretraining knowledge of a model (Zhang and Wu 2024; Gekhman et al. 2024; Kotha, Springer, and Raghunathan 2024; Wang 2024; Yang et al. 2024a; Li et al. 2025; Ye et al. 2025). Thus, a fine-tuned model may improve on the training task while losing out-of-domain factual reliability.

A decline in out-of-distribution (OOD) factual performance following SFT is frequently attributed to catastrophic forgetting or factual degradation (Luo et al. 2025; Li et al. 2024; Zhang and Wu 2024; Wu et al. 2025). However, reduced exact-match accuracy during open-ended generation does not conclusively demonstrate fact erasure. The model may merely lose access to the information, recognize correct answers only from candidates, or produce semantically correct responses that fail to satisfy strict evaluation criteria. Recent literature on spurious forgetting indicates that such performance decreases often reflect shifts in alignment or elicitation strategies rather than a genuine loss of stored knowledge (Zheng et al. 2025). Genuinely erased facts require restoration, whereas recognizable yet unreliable facts indicate failures in knowledge access or output expression.

To distinguish these possibilities, we compare classification-style evaluation with open-ended generation, following recent concerns that different answer formats probe different aspects of LLM knowledge and behavior (Tan et al. 2025; Chen et al. 2023; Elhady, Agirre, and Artetxe 2025; Rahmani et al. 2025). Classification-style probes test whether the model can select or rank the correct answer from candidates, whereas open-ended generation requires producing the answer directly. We find a clear dissociation after domain SFT: recognition-style performance remains comparatively stable on several OOD benchmarks, while open-ended factual generation declines sharply. A paired evaluation of the same facts under multiple-choice and open-ended formats further shows that many facts remain selectable but are no longer generated correctly. Failure-mode analysis reveals both genuinely incorrect answers and expression-level failures, including excessive verbosity, formatting mismatch, and exact-match artifacts (Wang et al. 2023, 2024; Li et al. 2025). We call this gap between retained factual capability and failed open-ended generation factual access failure, with expression failures as a major subclass.

Motivated by this diagnosis and by prior work on mitigating forgetting through pretraining-simulation, replay, or reference-model regularization (Chen et al. 2020; Yang et al. 2024b; Chen et al. 2026), we introduce Recall-Anchored Distillation (RAD), a base-anchored self-distillation objective

*These authors contributed equally.
†Corresponding author.

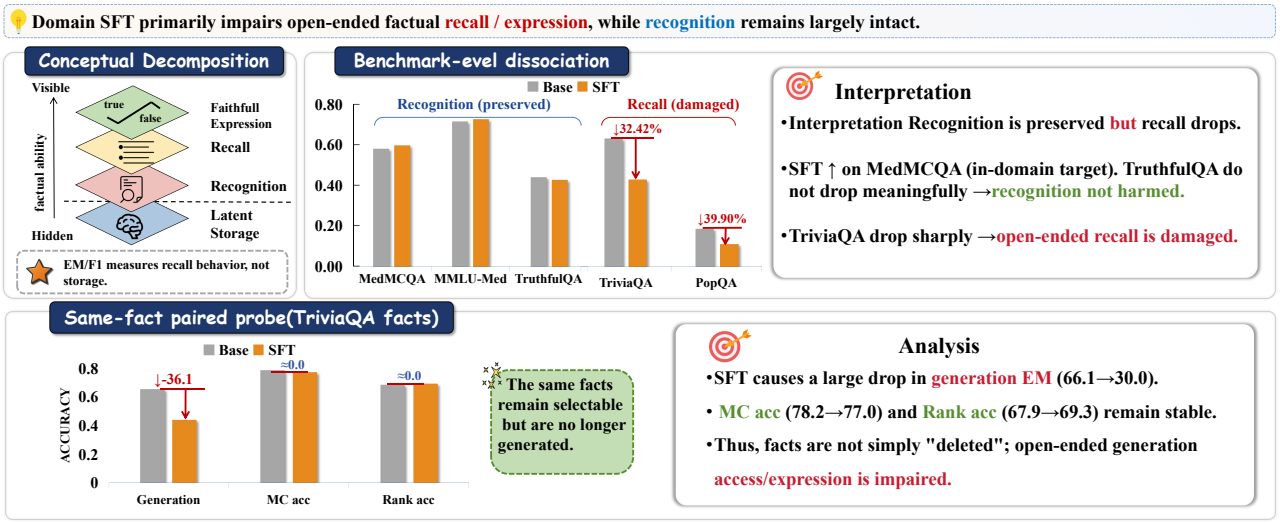

Figure 1: Diagnostic overview of factual access failures. Panel (a) separates latent factual storage from observable behaviors: recognition, recall, and expression. Panel (b) shows the benchmark-level comparison in Finding 1. Panel (c) shows the same-fact diagnosis in Finding 2.

for domain adaptation. RAD combines the standard supervised target-domain loss with an unlabeled OOD anchoring stream. For each OOD prefix–continuation example, the base model provides a soft next-token distribution, and the adapted model is trained to match it on continuation tokens. This distributional, rather than answer-based, anchor requires no gold OOD answers, rationales, external judges, or labeled factual data. Using the base model's continuation distribution as a reference (Hinton, Vinyals, and Dean 2015), RAD limits drift from pre-adaptation OOD generation behavior while allowing target-domain learning. Our contributions are as follows:

- We identify factual access failure as a diagnostic framework for SFT-induced degradation in open-ended OOD factual generation, supported by three complementary analyses.
- We propose Recall-Anchored Distillation (RAD), a base-anchored self-distillation method that preserves factual access by aligning to the base model's soft OOD continuation distribution.
- We evaluate RAD on three backbones fine-tuned on MedMCQA, showing that it recovers OOD factual generation while preserving in-domain gains and outperforming baselines.

Related Work

SFT-Induced Factual Degradation in LLMs

Recent work shows that such adaptation can unintentionally degrade factual reliability (Gekhman et al. 2024; Kaplan et al. 2026; Gong, Huang, and Liang 2025). In factual knowledge injection, Gekhman et al. show that knowledge unsupported by the base model is learned more slowly and can increase hallucinations once learned (Gekhman et al. 2024); Kang et al. show that unfamiliar fine-tuning examples can shape hallucinated prediction forms (Kang et al. 2024); and Zucchet et al. study factual-recall learning dynamics during fine-tuning (Zucchet et al. 2025). Complementary work on factual QA fine-tuning studies factually correct supervision, showing that fine-tuning can still alter the extraction of pretrained factual associations: models may over-amplify task-specific shortcuts, exhibit frequency shocks, or rely on relation-level cues while ignoring subject-specific information (Kazemi, Mittal, and Ramachandran 2023; Ghosal, Hashimoto, and Raghunathan 2024; Gong, Huang, and Liang 2025). Together, these studies show that SFT can degrade factual behavior even without noisy supervision. However, they leave open a finer-grained question: when closed-book factual accuracy drops after SFT, has the model lost the underlying fact, failed to access a still-recognizable fact in open-ended generation, or expressed the answer in a surface form that closed-book metrics cannot reliably match?

Hidden Knowledge and Recall Failures

Representation engineering studies how model behaviors and concepts are encoded in internal representations and how these representations can be analyzed or manipulated to control model outputs. Studies of implicit inference argue that fine-tuning can shift the model's inferred task distribution toward fine-tuning data, suppressing pretrained capabilities recoverable under alternative elicitation conditions (Kotha, Springer, and Raghunathan 2024). Work on spurious forgetting similarly shows that apparent forgetting can result from disrupted task alignment rather than true knowledge loss (Zheng et al. 2025). Complementary factual probing studies further show that models may encode more information than direct generation reveals, and that ranking-based, internal, or alternative probes can recover factual knowledge missed by standard closed-book generation (Gekhman et al. 2025; Orgad et al. 2025). These findings motivate our focus on factual

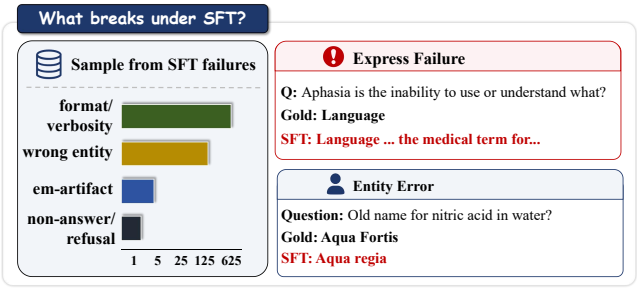

Figure 2: Failure modes among Base-correct/SFT-wrong examples. Most recall failures are format or verbosity mismatches rather than clean wrong-entity errors.

access. Rather than assuming that SFT-induced generation failures reflect storage loss, we ask whether the same facts remain recognizable under constrained probes and where extraction breaks down. Our repair strategy relates closely to knowledge distillation and KL-constrained training, which stabilize learning by matching a teacher or reference distribution (Hinton, Vinyals, and Dean 2015; Ziegler et al. 2020; Stiennon et al. 2022; Ouyang et al. 2022). Unlike generic KL constraints that keep an adapted policy close to a reference model, our method uses the base model as a teacher only on OOD factual continuations, namely general-domain factual continuations outside the SFT target domain, and regularizes the adapted model toward the base model's soft token distribution while optimizing target-domain SFT behavior.

Preliminary Studies

This section diagnoses factual degradation after domain SFT. A drop in closed-book generation accuracy shows failure to produce the correct answer, but not erasure of the corresponding fact. Consistent with recent studies on spurious forgetting and the distinction between knowledge encoding and recall (Calderon et al. 2026; Gekhman et al. 2025; Kotha, Springer, and Raghunathan 2024; Zheng et al. 2025), apparent degradation can reflect task-alignment shifts rather than loss of underlying capabilities. We distinguish memory erasure from expression failure through **three findings**.

F1: Models Recognize What They Cannot Recall

We compare recognition and recall at the benchmark level because domain SFT may broadly degrade factual behavior or selectively impair fact elicitation. Recognition tests whether the model selects correct answers under constrained evaluation, whereas recall tests whether it generates them without candidate support. We measure recognition with multiple-choice evaluation and recall with closed-book QA metrics, EM/F1 (Roberts, Raffel, and Shazeer 2020). Domain SFT improves the target-domain QA task, but this gain coincides with substantial degradation in open-ended factual recall on held-out QA benchmarks (see Figure 1b): for example, TriviaQA EM drops from 65.03 to 43.95 on the full benchmark, while MedMCQA and MMLU remain stable or slightly improve. This pattern is not yet a same-fact dissociation because datasets differ in content and format; instead, it motivates a controlled diagnosis of whether the same facts remain accessible under constrained evaluation.

F2: Facts Are Stored, but Inaccessible

To separate elicitation format from factual content, we evaluate recognition and recall on matched facts. We denote the factual association targeted by a QA item as $f = (s, r, o)$, where s, r, and o represent subject, relation, and answer object (Petroni et al. 2019; Ghosal, Hashimoto, and Raghunathan 2024); this is an analytic notation rather than a claim that TriviaQA items form explicit knowledge graphs. For each item, we pair the original open-ended question with recognition-style probes for the same answer object. On this subset, SFT substantially reduces open-ended generation (see Figure 1c): TriviaQA EM drops from 63.1 to 43.8, with marginal changes in multiple-choice accuracy and no ranking degradation. Under teacher forcing, the gold answer remains highly ranked: its first token stays near the top, with the top-1 rate nearly unchanged from Base to SFT. Thus, unreliably generated answers remain selectable or highly preferred under constrained or teacher-forced evaluation. We therefore use factual access failure as a behavioral diagnosis: the model selectively fails to produce certain facts in open-ended generation, rather than uniformly losing factual behavior.

F3: Recall Failures Are Largely Expression Failures

The paired probe shows that recognition and recall can diverge, motivating our analysis of these recall failures in generated text. We analyze 300 TriviaQA examples where the base model is EM-correct but the SFT model is EM-wrong. To distinguish factual errors from surface-form expression failures, we label each SFT output as a format/verbosity mismatch, wrong entity, EM artifact (semantically valid but rejected by exact match), or non-answer/refusal. The labels agree with an independent LLM judge (DeepSeek; In Appendix) on 97.3% of cases. As Figure 2 shows, format/verbosity mismatch dominates, accounting for 77.3% of the analyzed cases. This suggests that expression failures explain most Base-correct/SFT-wrong errors, although a non-negligible subset still corresponds to genuine wrong-answer generations. As a classifier-independent elicitation check, a simple 4-shot prompt recovers 90.0% of the same cases, indicating that many failures in this sample remain recoverable under different prompting rather than reflecting irreversible knowledge erasure. Together, these findings motivate a method that preserves open-ended factual expression during domain SFT.

Methodology

Motivated by the diagnosis in Section Preliminary, we introduce Recall-Anchored Distillation (RAD), a base-anchored self-distillation objective designed to preserve out-of-distribution (OOD) generation behavior during domain adaptation. Here, OOD is defined relative to the target domain. Figure 3 illustrates the overall RAD design.

Teacher and Student from One Model

Let f_{Θ_0} denote the original base model, which induces the next-token distribution $\pi_0(\cdot \mid u) = \mathrm{softmax}(f_{\Theta_0}(u))$

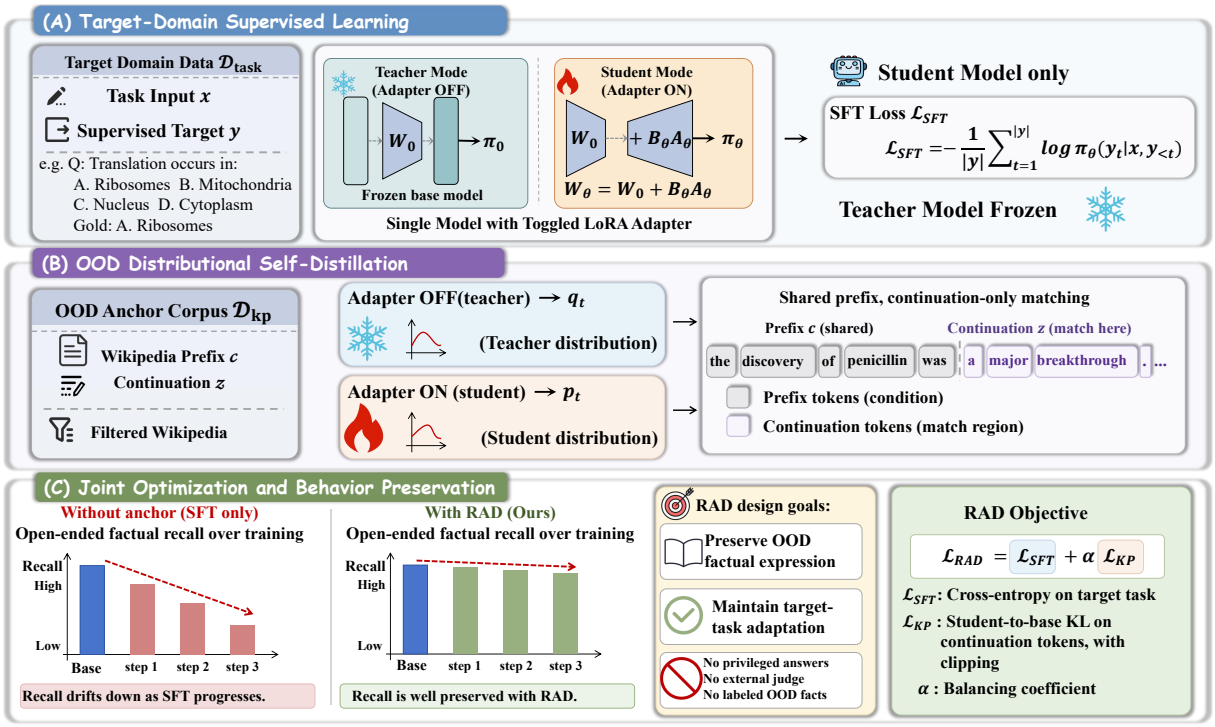

Figure 3: Overview of Recall-Anchored Distillation (RAD).

for input prefix u. We adapt it with Low-Rank Adaptation (LoRA) (Hu et al. 2021): each frozen weight matrix $W_0 \in \Theta_0$ receives a trainable low-rank update $W_\theta = W_0 + B_\theta A_\theta$, yielding the adapted student distribution $\pi_\theta(\cdot \mid u)$. RAD implements teacher and student in one model through the LoRA switch: adapter-off mode outputs the base distribution π_0 as the teacher, and adapter-on mode outputs π_θ as the student. Thus, the teacher is not an external oracle but the original base model itself, anchoring adaptation to the pre-adaptation generation behavior of the base model and following the broader use of teacher distributions to preserve model behavior during adaptation (Hinton, Vinyals, and Dean 2015; Li and Hoiem 2017).

Training Streams

Target-domain Supervision. The primary training signal comes from the supervised target-domain dataset $\mathcal{D}_{\text{task}} = \{(x, y)\}$, where x denotes the task input and y denotes the supervised target. This term is identical to standard domain SFT: it teaches the adapted model to solve the target-domain task and is the only component of RAD that uses task labels. For a mini-batch $\mathcal{B}_{\text{task}} \subset \mathcal{D}_{\text{task}}$, the adapter-on student is optimized with the teacher-forced cross-entropy objective:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{|\mathcal{B}_{\text{task}}|} \sum_{(x,y)\in\mathcal{B}_{\text{task}}} \frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}). \tag{1}$$

This loss drives domain adaptation by updating only LoRA parameters while keeping the base model frozen.

OOD Anchor Construction. The secondary training stream is an unlabeled OOD anchor corpus $\mathcal{D}_{\text{kp}} = \{(c, z)\}$, where c denotes a Wikipedia prefix and z its natural continuation. The continuation z is not treated as a hard label. Instead, it specifies the token positions over which the adapted model aligns with the base model next-token distribution, consistent with soft teacher distributions in knowledge distillation (Hinton, Vinyals, and Dean 2015). Since the diagnosis attributes the degradation primarily to OOD open-ended generation rather than target-domain learning itself, RAD sources this anchor stream from generic OOD text instead of target-domain training data.

To mitigate potential benchmark contamination, we filter the anchor corpus via case-insensitive substring matching, discarding any prefix–continuation pair that contains evaluation questions, answer strings, or normalized aliases from benchmarks including PopQA and TriviaQA. For each retained anchor example, we concatenate the prefix and continuation into a single sequence $s = [c; z]$. The prefix provides contextual grounding, while the continuation designates the matching region. We define a binary continuation mask $m_t \in \{0, 1\}$, where $m_t = 1$ indicates that the prediction target s_t falls within z, and $m_t = 0$ for prefix positions.

Distributional Anchoring Objective

RAD processes each OOD sequence in two model modes. With the adapter disabled, the adapter-off teacher provides the base next-token distribution:

$$q_t = \pi_0(\cdot \mid s_{<t}). \tag{2}$$

	In-Domain ↑	**OOD Recognition** (multiple choice) ↑			**OOD Recall** (closed-book) ↑			
	MedMCQA	MMLU-Med	MMLU-Other	TruthfulQA	TriviaQA		PopQA	
Method	Acc	Acc	Acc	MC2	EM	F1	EM	F1
Llama-3.1-8B								
Base	58.14	71.54	64.82	44.17	65.03	71.04	19.30	23.67
+ Standard SFT	59.43	72.20	**65.51**	**42.65**	43.95	56.59	11.60	16.03
+ Replay	59.53	71.90	64.88	41.48	32.87	47.80	3.40	8.06
+ RAD (Ours)	**60.05**	**72.57**	65.24	42.07	**51.62**	**62.16**	**15.20**	**19.63**
Qwen2.5-7B-Instruct								
Base	56.18	76.45	73.26	64.75	32.96	43.06	5.51	11.18
+ Standard SFT	61.27	76.04	**73.75**	49.03	23.27	**31.41**	4.95	8.77
+ Replay	61.68	77.06	73.47	50.21	16.09	25.31	1.77	6.11
+ RAD (Ours)	**61.99**	**77.18**	73.68	**58.14**	**23.68**	30.60	**6.81**	**9.99**
Qwen2.5-3B-Instruct								
Base	51.54	68.87	66.74	58.74	30.57	39.10	2.97	8.01
+ Standard SFT	**55.75**	67.97	67.45	45.89	1.04	13.58	1.45	5.17
+ Replay	55.22	**68.99**	67.15	44.99	3.39	13.24	0.98	5.38
+ RAD (Ours)	55.41	68.83	**68.12**	**54.88**	**9.04**	**22.00**	**2.66**	**6.77**

Table 1: Main results across three backbones on in-domain MedMCQA, OOD recognition, and OOD closed-book recall.

With the adapter enabled, the adapter-on student provides the adapted next-token distribution:

$$p_t = \pi_\theta(\cdot \mid s_{<t}). \tag{3}$$

RAD aligns the student to the frozen base distribution using reverse KL, with the optimized student distribution as the first argument and the base distribution as the reference:

$$d_t = D_{\mathrm{KL}}\left(p_t \parallel q_t\right). \tag{4}$$

This direction discourages the adapted model from assigning probability mass to OOD continuations that the base model deems improbable. To prevent a few highly divergent tokens from dominating the objective, we clip the per-token penalty at threshold τ. The OOD distillation loss is:

$$\mathcal{L}_{\mathrm{KP}} = \frac{1}{\sum_t m_t} \sum_t m_t \min(d_t, \tau). \tag{5}$$

The mask ensures that prefix tokens condition both model modes but contribute no gradient through the distillation loss.

Joint Optimization and Implementation

Joint Objective. RAD jointly optimizes target-domain adaptation and OOD distributional self-distillation via:

$$\mathcal{L}_{\mathrm{RAD}} = \mathcal{L}_{\mathrm{SFT}} + \alpha \mathcal{L}_{\mathrm{KP}}, \tag{6}$$

where α controls the strength of the OOD anchor. We set $\alpha = 1.5$ and $\tau = 5.0$ as default values. The supervised loss drives target-domain learning, while the OOD distillation loss discourages the adapted model from drifting away from the base model on generic continuations. RAD therefore does not freeze or constrain the adapter globally; it allows target-task adaptation while selectively anchoring the continuation distribution on OOD text.

Single-model Implementation. RAD is memory-efficient because it requires only one model instance in memory. For target-domain batches, we enable the LoRA adapter and compute $\mathcal{L}_{\mathrm{SFT}}$. For OOD anchor batches, we run the same sequence twice: first with the adapter disabled under `no_grad` to obtain teacher logits, and then with the adapter enabled to obtain student logits for computing $\mathcal{L}_{\mathrm{KP}}$. Only the adapter-on student pass contributes gradients. This implementation avoids loading a separate teacher model and makes the teacher-student distinction a mode switch of the same backbone. The additional cost is one extra forward pass per OOD anchor batch for teacher-logit computation. Since RAD pairs each target-domain optimizer step with one OOD anchor batch, the method adds modest training-time overhead while preserving the same inference-time architecture as standard LoRA SFT.

Experiments

Experimental Setup

Training Setup. We evaluate three backbones with different sizes and instruction-tuning properties: Llama-3.1-8B, Qwen2.5-7B-Instruct, and Qwen2.5-3B-Instruct (Qwen et al. 2025). For Llama-3.1-8B, we use the base checkpoint rather than the instruction-tuned variant. All adapted systems use LoRA adapters (Hu et al. 2021) with the pretrained backbone frozen, and all methods share the same target-domain data and evaluation protocol. RAD defaults to reverse KL ($\alpha = 1.5$, $\tau = 5.0$), with full details in Appendix.

Training Data. We use MedMCQA (Pal, Umapathi, and Sankarasubbu 2022) as the target-domain fine-tuning task, training adapted models on the training split and evaluating them on the validation split. Each example is a four-choice QA instance whose target response contains the correct option and its explanation. For the OOD anchor stream, RAD uses an unlabeled Wikipedia-style corpus, following the use

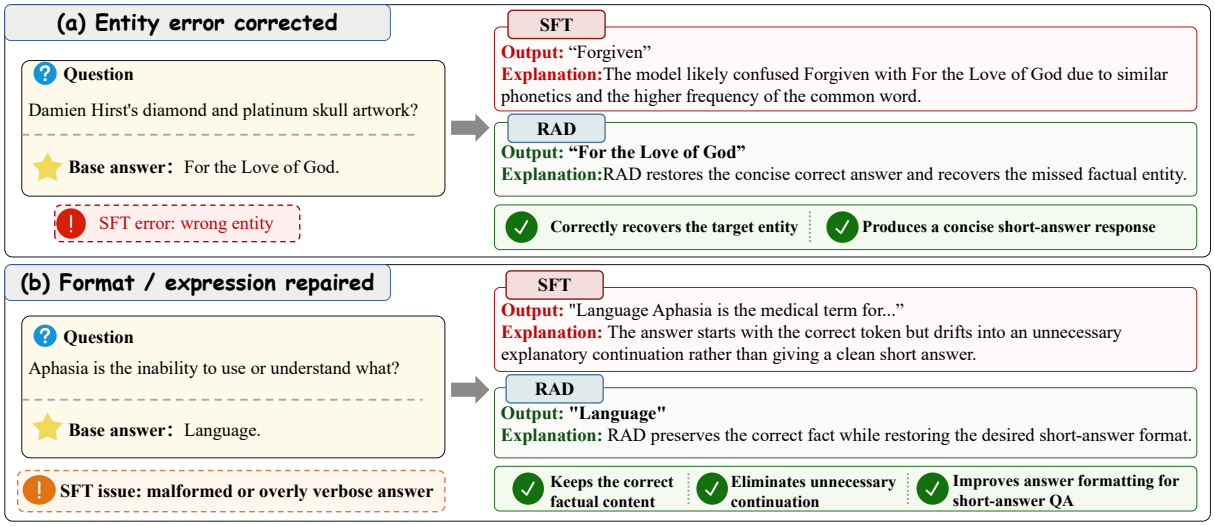

Figure 4: Qualitative examples showing how RAD repairs open-ended factual access failures on TriviaQA.

of Wikipedia-derived corpora in language modeling benchmarks (Merity et al. 2016). Here OOD is defined relative to the target-domain task. The anchor corpus contains 10,000 prefix–continuation examples, where the prefix provides context and the continuation marks the tokens for distributional anchoring. We filter this corpus against PopQA and TriviaQA entities and aliases to reduce benchmark contamination.

Compared Baselines. We compare RAD against three systems. **Base** is the original backbone without task adaptation and serves as a no-adaptation reference. **Standard SFT** fine-tunes the model only on MedMCQA, following the standard supervised fine-tuning paradigm used for adapting language models to target behaviors or tasks (Ouyang et al. 2022). **Replay** trains on the same amount of OOD anchor text as RAD, but uses a hard-label language-modeling loss rather than matching the base model's next-token distribution. This baseline controls for whether the benefit comes simply from adding general-domain text.

Benchmarks and Metrics

We evaluate target-domain adaptation on MedMCQA (Pal, Umapathi, and Sankarasubbu 2022) and report validation accuracy. To assess recognition-style OOD behavior, we use MMLU (Hendrycks et al. 2021), separating medical subjects as MMLU-Med and non-medical subjects as MMLU-Other, together with TruthfulQA-MC2 (Lin, Hilton, and Evans 2022). For open-ended OOD factual generation, we evaluate TriviaQA (Joshi et al. 2017) and PopQA (Mallen et al. 2023), reporting exact match (EM) and token-level F1 over acceptable aliases. All multiple-choice scoring, generation settings, normalization rules, and additional generation-based multiple-choice checks are detailed in Appendix.

Main Results

For Q1: Does Domain SFT Cause Capability Collapse or Selective OOD Shift? Table 1 shows domain SFT induces a selective OOD shift, not uniform capability degradation. Across three backbones, standard SFT improves in-domain MedMCQA performance, confirming effective target-domain adaptation. Recognition-style OOD performance remains comparable on MMLU-Med and MMLU-Other, suggesting domain tuning does not simply reduce general multiple-choice competence. Degradation is largest in open-ended or truthfulness-oriented OOD behavior. For `Llama-3.1-8B`, TriviaQA EM drops from 65.03 to 43.95, and for `Qwen2.5-3B-Instruct` from 30.57 to 1.04. `Qwen2.5-7B-Instruct` shows a related pattern, with TruthfulQA-MC2 dropping from 64.75 to 49.03. These results suggest that domain SFT primarily perturbs pre-adaptation knowledge expression under OOD conditions rather than uniformly erasing recognition-style ability.

For Q2: Can RAD Recover OOD Factual Behavior While Preserving Target-Domain Adaptation? RAD mitigates the SFT-induced OOD shift while retaining competitive in-domain performance. On `Llama-3.1-8B`, RAD improves both TriviaQA and PopQA EM, while achieving the best MedMCQA accuracy among fine-tuned systems. The same trend extends to the Qwen backbones, though the affected OOD axis differs by model: RAD substantially recovers TruthfulQA-MC2 on `Qwen2.5-7B` and TriviaQA EM on `Qwen2.5-3B`, with only a small MedMCQA drop for the latter. These results suggest that anchoring the adapted model to the base model on unlabeled OOD continuations can partly recover damaged generation behavior without materially undermining target-domain learning. Since RAD introduces no annotated OOD facts, the improvement is better viewed as preservation of generation behavior rather than re-learning the evaluation answers.

For Q3: Is the Soft Base Distribution Necessary Beyond Same-Text Replay? Replay isolates the role of the soft base distribution: it uses the same OOD anchor corpus as RAD but replaces distributional anchoring with hard-label language modeling. If RAD gained only from extra general-

Variant	MedMCQA	TriviaQA EM	PopQA EM
Standard SFT	59.43	43.95	11.60
RAD default	60.05	51.62	15.20
Forward KL	59.89	52.88	16.74
Symmetric KL	59.93	52.57	16.58
No clipping	59.98	51.59	16.58
$\tau = 2.5$	59.98	52.17	16.77
$\tau = 10.0$	60.12	51.84	16.60
On-policy anchor	60.24	51.44	15.10

Table 2: Ablation results of RAD.

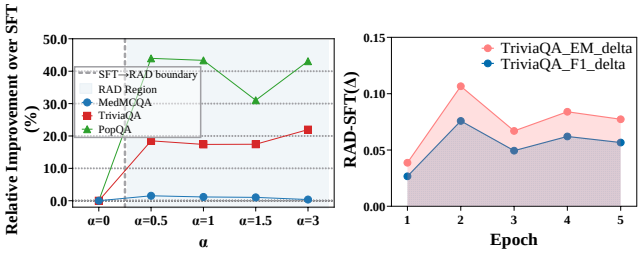

Figure 5: Ablation and checkpoint-dynamics results on `Llama-3.1-8B`. Left: effect of anchoring strength α. Right: TriviaQA accuracy gains of RAD over Standard SFT at each training epoch.

domain text, Replay would provide comparable preservation. Instead, Replay lags RAD on OOD recall and often worsens SFT degradation. This is clearest for `Llama-3.1-8B`, where Replay reduces PopQA EM to 3.40, versus 11.60 under standard SFT and 15.20 under RAD. Thus, the key signal is not anchor text alone but the token-level base-model distribution over that text. Matching this soft distribution preserves relative preferences over OOD continuations, whereas hard-label replay adds a competing objective that does not directly protect generation behavior disrupted by domain SFT.

Ablation Analysis. We conduct an ablation study of RAD on Llama-3.1-8B to evaluate the impact of anchoring strength, checkpoint dynamics, KL divergence direction, clipping thresholds, and sampling strategies. Figure 5 demonstrates that the performance improvements are driven by positive OOD anchoring. Compared to the standard SFT baseline ($\alpha = 0$), setting $\alpha > 0$ enhances performance on TriviaQA and PopQA while maintaining accuracy on MedMCQA near the baseline level. The observed non-monotonic gains indicate that the parameter α regulates the degree of OOD preservation rather than converging to a universal optimum; consequently, we adopt $\alpha = 1.5$ as a stable default configuration. Across all five training epochs, RAD consistently yields higher accuracy on TriviaQA compared to standard SFT, indicating that the benefits are not specific to the final checkpoint. Furthermore, Table 2 illustrates the robustness of the method across variations in forward KL, symmetric KL, alternative clipping thresholds, and on-policy anchoring, with EM scores exceeding 51% on TriviaQA and remaining near 60% on MedMCQA. These findings suggest that performance preservation stems from aligning the adapted model with the soft OOD continuation distribution of the base model, rather than merely from exposure to the anchor text.

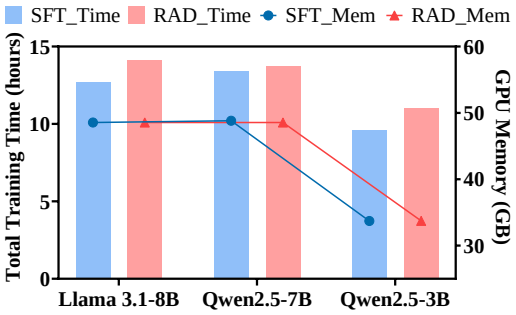

Figure 6: Training time and GPU memory usage of Standard SFT and RAD across three backbones.

Qualitative Analysis. Figure 4 illustrates qualitative examples of failure modes identified in the preliminary diagnosis. In the entity-error case, SFT replaces target answer with an incorrect but plausible entity, whereas RAD recovers the concise factual answer. In the format-error case, SFT starts with the correct token but continues into an unnecessary explanatory completion, producing a malformed short answer; RAD preserves the factual content and restores the desired answer format. These examples serve only as qualitative illustrations; aggregate trends appear in Table 1, Table 2 and Figure 5.

Efficiency Analysis. We examine RAD training cost relative to Standard SFT in Figure 6. Because RAD uses the same backbone in two LoRA modes instead of a separate teacher model, its memory footprint remains essentially unchanged across all three backbones, with differences within 1 GB. The main overhead is the additional adapter-off forward pass on OOD anchor batches, but the total training-time increase is modest: the largest increase is about 1.4 hours. Overall, RAD adds limited training-time cost while preserving the same inference-time architecture as standard LoRA SFT.

Conclusion

This work studies factual access failure in domain SFT, where models recognize facts but fail to reliably recall or express them in open-ended OOD generation. Via benchmark comparisons, same-fact probes, and failure-mode analysis, we show SFT-induced degradation is selective: recognition remains stable while open-ended recall and expression degrade. We introduce Recall-Anchored Distillation (RAD), preserving OOD generation by aligning the adapted model with the base model's soft continuation distribution on unlabeled OOD text. Across three MedMCQA-tuned backbones, RAD partially recovers damaged OOD factual behavior while maintaining target-domain adaptation and outperforming replay on identical anchor text.

References

Calderon, N.; Ben-David, E.; Gekhman, Z.; Ofek, E.; and Yona, G. 2026. Empty Shelves or Lost Keys? Recall Is the Bottleneck for Parametric Factuality. arXiv:2602.14080.

Chen, L.; Deng, Y.; Bian, Y.; Qin, Z.; Wu, B.; Chua, T.-S.; and Wong, K.-F. 2023. Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators. arXiv:2310.07289.

Chen, S.; Hou, Y.; Cui, Y.; Che, W.; Liu, T.; and Yu, X. 2020. Recall and Learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting. arXiv:2004.12651.

Chen, Y.; Zhu, T.; Zhang, M.; Chen, X.; Huang, J.; Xu, D.; and Wang, Y. 2026. On-Policy Replay for Continual Supervised Fine-Tuning. arXiv:2605.29495.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416.

Elhady, A.; Agirre, E.; and Artetxe, M. 2025. WiCkeD: A Simple Method to Make Multiple Choice Benchmarks More Challenging. arXiv:2502.18316.

Gekhman, Z.; David, E. B.; Orgad, H.; Ofek, E.; Belinkov, Y.; Szpektor, I.; Herzig, J.; and Reichart, R. 2025. Inside-Out: Hidden Factual Knowledge in LLMs. arXiv:2503.15299.

Gekhman, Z.; Yona, G.; Aharoni, R.; Eyal, M.; Feder, A.; Reichart, R.; and Herzig, J. 2024. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? arXiv:2405.05904.

Ghosal, G.; Hashimoto, T.; and Raghunathan, A. 2024. Understanding Finetuning for Factual Knowledge Extraction. arXiv:2406.14785.

Gong, X.; Huang, H.; and Liang, S. 2025. From Parameters to Prompts: Understanding and Mitigating the Factuality Gap between Fine-Tuned LLMs. arXiv:2505.23410.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. arXiv:1705.03551.

Kang, K.; Wallace, E.; Tomlin, C.; Kumar, A.; and Levine, S. 2024. Unfamiliar Finetuning Examples Control How Language Models Hallucinate. arXiv:2403.05612.

Kaplan, G.; Gekhman, Z.; Zhu, Z.; Rozner, L.; Reif, Y.; Swayamdipta, S.; Hoiem, D.; and Schwartz, R. 2026. Why Fine-Tuning Encourages Hallucinations and How to Fix It. arXiv:2604.15574.

Kazemi, M.; Mittal, S.; and Ramachandran, D. 2023. Understanding Finetuning for Factual Knowledge Extraction from Language Models. arXiv:2301.11293.

Kotha, S.; Springer, J. M.; and Raghunathan, A. 2024. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. arXiv:2309.10105.

Li, H.; Ding, L.; Fang, M.; and Tao, D. 2024. Revisiting Catastrophic Forgetting in Large Language Model Tuning. arXiv:2406.04836.

Li, R.; Wang, X.; and Yu, H. 2024. LlamaCare: An Instruction Fine-Tuned Large Language Model for Clinical NLP. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10632–10641. Torino, Italia: ELRA and ICCL.

Li, Z.; Chen, C.; Xu, T.; Qin, Z.; Xiao, J.; Luo, Z.-Q.; and Sun, R. 2025. Preserving Diversity in Supervised Fine-Tuning of Large Language Models. arXiv:2408.16673.

Li, Z.; and Hoiem, D. 2017. Learning without Forgetting. arXiv:1606.09282.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958.

Luo, Y.; Yang, Z.; Meng, F.; Li, Y.; Zhou, J.; and Zhang, Y. 2025. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. arXiv:2308.08747.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. arXiv:2212.10511.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer Sentinel Mixture Models. arXiv:1609.07843.

Orgad, H.; Toker, M.; Gekhman, Z.; Reichart, R.; Szpektor, I.; Kotek, H.; and Belinkov, Y. 2025. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. arXiv:2410.02707.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. arXiv:2203.14371.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. Hong Kong, China: Association for Computational Linguistics.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Rahmani, H. A.; Krishna, S.; Wang, X.; Naghiaei, M.; and Yilmaz, E. 2025. Self-Correcting Large Language Models: Generation vs. Multiple Choice. arXiv:2511.09381.

Roberts, A.; Raffel, C.; and Shazeer, N. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5418–5426. Online: Association for Computational Linguistics.

Savage, T.; Ma, S.; Boukil, A.; Patel, V.; Rangan, E.; Lopez, I.; and Chen, J. H. 2024. Fine Tuning Large Language Models for Medicine: The Role and Importance of Direct Preference Optimization. arXiv:2409.12741.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; Schaekermann, M.; Wang, A.; Amin, M.; Lachgar, S.; Mansfield, P.; Prakash, S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Tomasev, N.; Liu, Y.; Wong, R.; Semturs, C.; Mahdavi, S. S.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. 2022. Learning to summarize from human feedback. arXiv:2009.01325.

Tan, C.; Shao, W.; Xiong, H.; Zhu, T.; Liu, Z.; Shi, K.; and Chen, W. 2025. UAQFact: Evaluating Factual Knowledge Utilization of LLMs on Unanswerable Questions. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 1700–1715. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Wang, C.; Cheng, S.; Guo, Q.; Yue, Y.; Ding, B.; Xu, Z.; Wang, Y.; Hu, X.; Zhang, Z.; and Zhang, Y. 2023. Evaluating Open-QA Evaluation. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 77013–77042. Curran Associates, Inc.

Wang, S. 2024. Real-Time Object Detection Using a Lightweight Two-Stage Detection Network with Efficient Data Representation. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 1(1): 17–30.

Wang, Y.; Wang, M.; Manzoor, M. A.; Liu, F.; Georgiev, G.; Das, R. J.; and Nakov, P. 2024. Factuality of Large Language Models: A Survey. arXiv:2402.02420.

Wu, C.-C.; Tam, Z. R.; Lin, C.-Y.; Chen, Y.-N.; Sun, S.-H.; and yi Lee, H. 2025. Mitigating Forgetting in LLM Fine-Tuning via Low-Perplexity Token Learning. arXiv:2501.14315.

Yang, H.; Zhang, Y.; Xu, J.; Lu, H.; Heng, P. A.; and Lam, W. 2024a. Unveiling the Generalization Power of Fine-Tuned Large Language Models. arXiv:2403.09162.

Yang, X.; Huang, J.; Li, J.; Zhao, Y.; Li, H.; Yu, Z.; Gao, S.; and Cao, R. 2024b. Optically Mediated Nonvolatile Resistive Memory Device Based on Metal–Organic Frameworks. *Advanced Materials*, 36(35): 2313608.

Ye, J.; Yang, Y.; Nan, Y.; Li, S.; Zhang, Q.; Gui, T.; Huang, X.; Wang, P.; Shi, Z.; and Fan, J. 2025. Analyzing the Effects of Supervised Fine-Tuning on Model Knowledge from Token and Parameter Levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 471–513. Association for Computational Linguistics.

Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; and Wang, G. 2025. Instruction Tuning for Large Language Models: A Survey. arXiv:2308.10792.

Zhang, X.; and Wu, J. 2024. Dissecting Learning and Forgetting in Language Model Finetuning. In *International Conference on Learning Representations*.

Zheng, J.; Cai, X.; Qiu, S.; and Ma, Q. 2025. Spurious Forgetting in Continual Learning of Language Models. arXiv:2501.13453.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593.

Zucchet, N.; Bornschein, J.; Chan, S.; Lampinen, A.; Pascanu, R.; and De, S. 2025. How do language models learn facts? Dynamics, curricula and hallucinations. arXiv:2503.21676.