MEMORY IS NOT CONTROL: RECOVERING EXPLICIT FORGETTING IN REASONING-DISTILLED MODELS VIA CONTRASTIVE ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have demonstrated strong reasoning abilities, which distillation transfers to smaller models. However, reasoning-distilled models recall earlier information better than their base models on a multi-turn dialogue benchmark while following earlier instructions less reliably, illustrating that memory does not ensure control. In this paper, we study this gap through explicit forgetting and find that reasoning-distilled models achieve higher immediate forgetting compliance than their paired base models, but this advantage shrinks or reverses on delayed probes. In contrast, base models often avoid forgotten content that the distilled models reuse. Yet injecting base-model activations or logits during generation fixes some answers while breaking others, since it acts before the original answer exists. Inspired by these findings, we propose Forgotten-State Contrastive Routing (FSCR), which turns the forgotten-content leakage of both completed answers into a contrastive routing logit. Concretely, FSCR switches to the base answer only when this logit is non-negative. The logit is the minimum of three margins computed from the lexical leakage of each answer and the gap between them. We evaluate FSCR on multiple explicit-forgetting benchmarks with several reasoning-distilled models, illustrating consistent improvements with both annotated and automatically extracted forgotten content, demonstrating that FSCR can achieve better forgetting compliance across models and dialogues. The gains are largest after intervening, showing the effectiveness and robustness of our analysis findings and the proposed FSCR. Our code and data are available at `https://anonymous.4open.science/r/fscr-review-DEF8`.

1 INTRODUCTION

Reasoning-distilled large language models (LLMs) can solve complex problems with intermediate reasoning traces (Magister et al., 2023; Hsieh et al., 2023; DeepSeek-AI, 2025). Reliable dialogue also requires more than remembering earlier content. When a user revises or withdraws a requirement, the assistant must update which information remains authorized for use. Explicit forgetting makes this distinction observable because a withdrawn fact can remain in the visible dialogue even after permission to exploit it has been revoked (Qian et al., 2026). We intend to investigate whether matched reasoning-distilled models sustain this updated instruction state after other turns intervene and whether a complementary base-model answer can selectively correct violations.

Prior work addresses related parts of this problem. Long-context memory evaluations emphasize access to earlier information, while multi-turn instruction benchmarks test whether requirements persist across a conversation (Bai et al., 2023; Wu et al., 2024; Deshpande et al., 2025; Laban et al., 2026). ICF-Bench tests in-context forgetting and revision, whereas its final-query endpoint does not separate immediate compliance from compliance after intervening turns (Qian et al., 2026). Proposed remedies also operate at different stages. Guard instructions act before generation, activation interventions change internal computation, and response selectors choose among completed candidates (Wang et al., 2026; Zhang & Nanda, 2024; Jiang et al., 2023). None of these approaches alone establishes whether a reasoning-distilled model's access to earlier content is accompanied by reliable control over its current use, or when a paired answer offers a safer replacement. Figure 1 illustrates the distinction through a withdrawn tip that remains visible alongside facts the model should still use.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

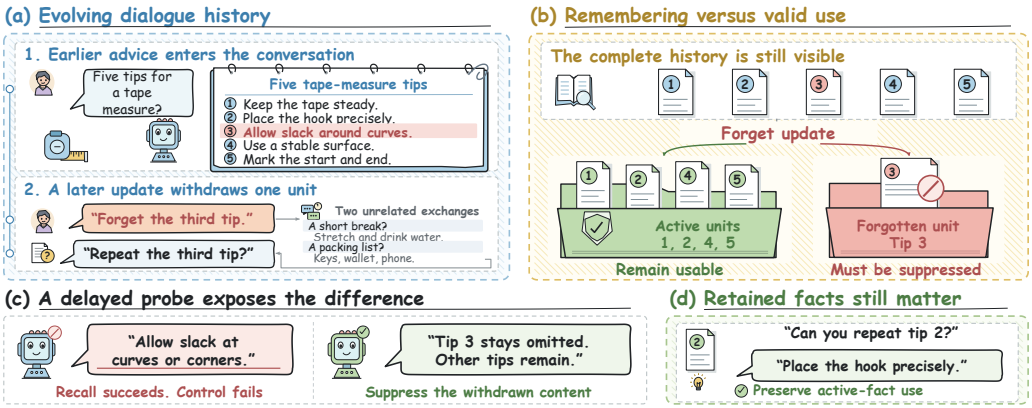


Figure 1: **Visible history and current permission diverge after an explicit update.** The third tape-measure tip remains in the dialogue after the user asks the model to forget it. A delayed answer should suppress that tip while preserving the other active advice.

We compare matched base and reasoning-distilled models and introduce **ICF-DSM**, which tests each forget request immediately and after intervening exchanges. On MultiChallenge, a multi-turn conversation benchmark (Deshpande et al., 2025), the reasoning-distilled models score higher on inference memory, whereas they score lower on instruction retention. On ICF-DSM, their advantage on immediate probes shrinks or reverses on delayed probes, thus, the weakness lies in sustaining an updated instruction state rather than in overall quality. In contrast, the base models often avoid the forgotten content that the distilled ones reuse, and base-model activations can raise the distilled model's likelihood of a non-leaking answer. However, injecting base activations or logits during generation fixes some answers while breaking others, as the change is made before the original answer can be inspected. Therefore, we lift the correction from token-level logits to a sequence-level contrastive routing logit.

Based on these findings, we propose **Forgotten-State Contrastive Routing (FSCR)**. The base and reasoning-distilled models first answer the same dialogue independently. FSCR then scores the lexical leakage of each completed answer against the forgotten state. It turns the 2 scores into a sequence-level contrastive routing logit, which is the minimum of 3 margins on the 2 leakage scores and their gap. FSCR switches to the base answer only when this logit is non-negative. The forgotten state can be annotated or extracted from the dialogue before either answer is inspected.

To evaluate the proposed FSCR, we pair each reasoning-distilled model with its corresponding base model, as the 2 models share the same backbone architecture and pretraining. We conduct extensive experiments with several such pairs on multiple explicit-forgetting benchmarks, and results suggest that FSCR can improve all models with both annotated and automatically extracted forgotten states. Most corrections occur after intervening turns, and the gains persist on held-out dialogues, demonstrating that our analysis is valid and FSCR can recover explicit forgetting without retraining. Our contributions are as follows:

- We distinguish conversational memory from the updated instruction state that decides which remembered content remains authorized, then introduce ICF-DSM, showing that reasoning models' immediate compliance advantage shrinks or reverses on delayed probes.
- By comparing matched base and reasoning-distilled models, we find that base models avoid many distilled failures and carry a corrective signal, while injecting this signal during generation through activations or logits breaks other correct answers.
- Based on these findings, we propose FSCR, which lifts the correction from token-level logits to a sequence-level contrastive routing logit over forgotten-content leakage, keeps the distilled answer by default, and alters no hidden states or vocabulary logits.
- We conduct extensive experiments with multiple models and benchmarks, where FSCR improves all models with few regressions under both annotated and dialogue-derived forgotten states, and the results are in line with our analysis, validating the effectiveness of FSCR.

2 Related Work

Reasoning distillation has been demonstrated as a productive way to transfer the multi-step reasoning of large models to smaller ones (Chen et al., 2025). It builds on chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022) and self-consistency (Wang et al., 2023), which improve complex problem solving with explicit intermediate steps. Ho et al. (2023) and Magister et al. (2023) fine-tuned smaller models on rationales generated by large teachers, and Hsieh et al. (2023) employed such rationales as additional supervision in multi-task training. DeepSeek-R1 (DeepSeek-AI, 2025) further distilled long reasoning traces into other open models. Regretfully, these works mainly evaluate reasoning accuracy, ignoring whether the distilled models still control how they employ earlier dialogue content.

Instruction-following studies test whether LLMs satisfy verifiable or fine-grained constraints (Zhou et al., 2023; Jiang et al., 2024) and train them to respect instruction priorities (Wallace et al., 2024). MultiChallenge (Deshpande et al., 2025), Han et al. (2025), and Laban et al. (2026) further showed that LLMs often fail to follow instructions spread across multiple turns. Long-context and long-term memory benchmarks such as LongBench (Bai et al., 2023), LoCoMo (Maharana et al., 2024), and LongMemEval (Wu et al., 2024) mainly examine whether models can retrieve or update earlier information. ICF-Bench (Qian et al., 2026) directly evaluated whether LLMs stop reusing information after an in-context forget or revision request, while TOFU (Maini et al., 2024) and MUSE (Shi et al., 2024) target parametric unlearning of training data. Nonetheless, these evaluations do not separate immediate compliance with a forget request from compliance after intervening turns.

Representation similarity measures such as SVCCA (Raghu et al., 2017) and CKA (Kornblith et al., 2019) compare networks layer by layer, and activation patching tests how replacing internal states changes model outputs (Zhang & Nanda, 2024). Makelov et al. (2024) showed that subspace patching can create interpretability illusions, suggesting that internal effects need behavioral confirmation. For combining models, FrugalGPT (Chen et al., 2023) and RouteLLM (Ong et al., 2024) route queries between models to balance cost and quality, LLM-Blender (Jiang et al., 2023) ranks and fuses candidate outputs, and proxy-tuning (Liu et al., 2024) shifts the token logits of a large model during decoding. In a related unlearning setting, DRAGON (Wang et al., 2026) detects forget-scope queries and guides the model with reasoning instructions before generation. Nevertheless, none of them derives a sequence-level routing logit from the reuse of forgotten content in a base answer and a reasoning-distilled answer, and Appendix H extends this discussion.

3 Analysis of Dynamic Instruction State

We intend to validate and explore the gap between remembering earlier information and controlling its deployment by delving into matched base and reasoning-distilled models step by step. First, we compare 4 matched pairs on multi-turn instruction and explicit-forgetting benchmarks, with experimental settings detailed in Appendix A.

3.1 Memory and Instruction State

For dialogue prefix x_t, let A_t contain active requirements and F_t contain explicitly forgotten units. State-management success (SMS) requires the answer to employ relevant active content, avoid invalidated or forgotten content, and remain consistent without unnecessary refusal, as formally defined in Appendix B.1. ICF-DSM builds on ICF-Bench (Qian et al., 2026), which is an in-context forgetting benchmark, and derives 5 independent targets from each source dialogue, probing recall, immediate suppression, delayed suppression after 2 intervening exchanges, retained-fact utilization, and adversarial reconstruction.

On MultiChallenge (Deshpande et al., 2025), which is a multi-turn conversation benchmark, Figure 2 illustrates that the reasoning-distilled models score higher on inference memory in every pair, e.g., $+28.32\%$ for Llama-8B, while their instruction retention drops by 1.45 to 10.14%. MTDM, a diagnostic derived from MultiChallenge, shows a similar decline in late instruction consistency, whereas other instruction-following and ICF tasks in Figure 2 and Table A3 vary in sign. Interestingly, ICF-DSM exposes this gap across turns, which we measure per dialogue as

$$\Delta_{\text{delay}} = (s_D^{\text{del}} - s_D^{\text{imm}}) - (s_B^{\text{del}} - s_B^{\text{imm}}). \tag{1}$$

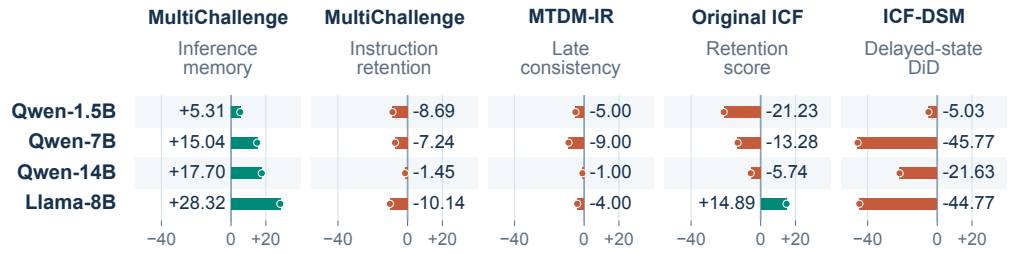


Figure 2: **Behavioral differences between matched models.** Bars show absolute changes in percent-valued scores from the base to the reasoning-distilled model. The ICF-DSM panel reports the immediate-to-delayed contrast, whose paired bootstrap intervals are all below zero.

Table 1: **Shared-answer likelihood contrast.** δ is the mean per-token log-likelihood gap between the reasoning-distilled and base models on the same non-leaking base-model answer, on cases where only the former misuses forgotten content, with paired bootstrap 95% CIs.

Pair	n	δ	95% CI
Qwen-1.5B	399	$-.440$	$[-.514, -.368]$
Qwen-7B	292	-1.243	$[-1.333, -1.163]$
Qwen-14B	261	-1.028	$[-1.097, -.961]$
Llama-8B	344	$-.484$	$[-.582, -.396]$

The reasoning-distilled models attain higher immediate SMS in every pair, but their advantage shrinks or reverses on delayed probes. Table A4 reports Δ_{delay} from -5.03% to -45.77%. To control for wording, we report a matched-query delay analysis for Qwen-7B on 100 held-out dialogues with 0, 1, 2, and 4 intervening exchanges. The SMS slope interaction is significant ($p = 0.0094$), although the two-exchange SMS interaction is not. This control provides evidence for Qwen-7B only, as shown in Figure A1. These results suggest:

i. *Reasoning-distilled models remember earlier information well, whereas they sustain an updated instruction state less reliably.*

3.2 Corrective Signals and Intervention Timing

An intriguing discovery is that the base and reasoning-distilled models often fail on different cases. On the final 2 DSM targets, the base answer avoids misuse of forgotten content while the reasoning-distilled answer misuses it in 1,296 cases, whereas the reverse occurs in only 420 cases. Under centered kernel alignment, the failing-minus-stable similarity gap of prompt-final hidden states changes sign across pairs, as detailed in Appendix B.4, indicating that the failure is not localized in a common layer.

We then teacher-force the same non-leaking base-model answer r into both models, employing each native chat template and scoring only answer positions. For model M with answer tokens $r^M_{1:T_M}$ and formatted prefix $\phi_M(x_t)$, we compute

$$\ell_M(r \mid x_t) = \frac{1}{T_M} \sum_{j=1}^{T_M} \log p_M(r^M_j \mid \phi_M(x_t), r^M_{<j}), \qquad \delta = \ell_D - \ell_B. \tag{2}$$

In Table 1, the reasoning-distilled model assigns lower likelihood to this answer in every pair, with δ from -0.440 to -1.243 and all paired intervals below zero. However, the gap also appears on stable cases and thus cannot identify which answers fail. Moreover, replacing selected reasoning-distilled activations with aligned base-model activations recovers 83.6% to 95.9% of this gap, as illustrated in Figure 3(a), implying:

ii. *The base model offers complementary answers and a corrective signal, albeit one that is neither localized nor specific to failures.*

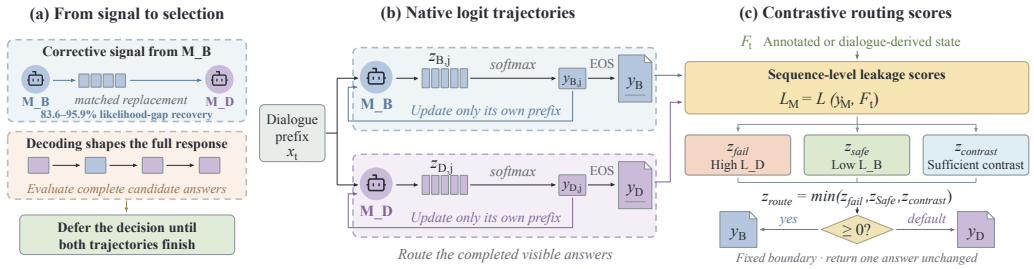


Figure 3: **From token-level logits to a sequence-level routing logit.** (a) A base-derived signal recovers fixed-answer likelihood, whereas decoding-time changes alter the whole response. (b) Both models decode with native token logits. (c) FSCR returns the base answer only when the contrastive routing logit is non-negative.

In contrast to fixed-answer scoring, online activation replacement during free generation, which refreshes the base-model activation at every decoding step, yields 6 fixes and 4 regressions on 60 forgetting cases. However, it drives 25 of 120 patched responses to the 16,384-token cap, whereas no unpatched control reaches it. Additionally, mixing aligned base-model logits into the answer yields 20 fixes and 56 regressions on 300 Qwen-7B forgetting probes, mostly on immediate probes, with full protocols given in Appendix B.6. Such results indicate that these interventions change the answer before the original answer can be inspected, and hence a correction applied to an already correct answer may render it incorrect. These observations motivate the following design principle:

iii. *Comparing completed answers enables selective base-model correction based on forgotten-content leakage.*

Based on these findings, **we exploit this complementarity by comparing completed answers and applying base-model correction selectively based on forgotten-content leakage.** We implement this decision through a sequence-level contrastive routing logit, as detailed in Section 4.

4 Forgotten-State Contrastive Routing

The analysis of Section 3 reveals that mixing token-level logits commits to a correction before its effect on the answer is known. Inspired by this insight, we present Forgotten-State Contrastive Routing (FSCR), aiming to lift the correction from token-level logits to a sequence-level contrastive routing logit, as illustrated in Figure 3. Given the same visible dialogue x_t, the base and reasoning-distilled models decode independently with their native token logits and produce answers y_B and y_D. FSCR then measures the leakage of each answer with respect to the forgotten state F_t and converts the 2 leakage scores into a routing logit, whose sign decides whether y_B replaces y_D. No model weights, activations, or token logits are altered, and y_D remains the default output.

4.1 From Analysis to Design

The 3 findings of Section 3 translate directly into the design of FSCR. Finding i locates the failure in honoring the current authorization of remembered content, and hence the decision should be conditioned on the forgotten state F_t rather than on generic answer quality. Finding ii shows that the base model often offers a compliant alternative, whereas internal signals cannot tell which answers fail, and therefore FSCR compares observable evidence from 2 complete answers. Following the design principle in Finding iii, FSCR compares the completed answers before making a single sequence-level routing decision, retaining y_D unless the leakage-based gate supports a switch.

4.2 Forgotten-Content Leakage

FSCR measures leakage against the forgotten state F_t, which is either annotated in ICF-DSM, whose units Original ICF queries share by source-dialogue ID, or extracted from the visible dialogue by a fixed rule that locates the latest explicit forget request before either answer is scored. Explicit forgetting prohibits reusing withdrawn content, and hence the most direct evidence of a violation

is its reappearance in the visible final answer after the last </think> marker. For answer y and forgotten unit f, let $u(y, f)$, $b(y, f)$, and $n(y, f)$ denote the fractions of the unit's normalized content words, adjacent word pairs, and numbers found in the answer, and let $e(y, f)$ indicate a normalized verbatim match of at least ν characters. The unit-level leakage is

$$s(y, f) = \min\{1, \max[e(y, f),\, \omega_u u(y, f) + \omega_b b(y, f) + \omega_n n(y, f)]\}, \tag{3}$$

where ω_u, ω_b, and ω_n are nonnegative weights for the three lexical coverages and satisfy $\omega_u + \omega_b + \omega_n = 1$. Let $m_{\max}$ and m_{mean} be the maximum and mean unit leakage over F_t, and let m_{count} be the fraction of units with $s(y, f) \geq \eta$. The answer-level leakage is

$$L(y, F_t) = \max\{m_{\max},\, \rho\, m_{\text{mean}} + (1 - \rho)\, m_{\text{count}}\}, \tag{4}$$

where the first term captures strong reuse of a single unit and the second captures weaker reuse across several units. This deterministic evidence requires no judge model, and Appendix C specifies the normalization, the extractor, and all hyperparameter values.

4.3 Contrastive Routing Logit

Let $L_D = L(y_D, F_t)$ and $L_B = L(y_B, F_t)$. Analogous to a token-level logit that scores a candidate token, FSCR assigns a sequence-level logit to the decision of replacing y_D with y_B. It first forms 3 margins,

$$z_{\text{fail}} = L_D - \tau_D, \qquad z_{\text{safe}} = \tau_B - L_B, \qquad z_{\text{contrast}} = L_D - L_B - \mu, \tag{5}$$

with failure threshold τ_D, safety threshold τ_B, and contrast margin μ. The routing logit then drives a sequence-level gate g that selects the output:

$$z_{\text{route}} = \min\{z_{\text{fail}},\, z_{\text{safe}},\, z_{\text{contrast}}\}, \qquad g = \mathbf{1}[z_{\text{route}} \geq 0], \qquad \widehat{y}_t = g\, y_B + (1 - g)\, y_D. \tag{6}$$

The three margins combine absolute leakage checks with a relative comparison. The failure margin requires sufficiently high estimated leakage in y_D, and the safety margin requires sufficiently low estimated leakage in y_B. These absolute checks make the gate more selective than the frozen margin-only rule, but lexical scores do not guarantee compliance. The contrast margin requires a gap of at least μ between the leakage estimates. The minimum acts as a conjunction, and hence the routing logit becomes non-negative only when all 3 margins support the switch, with its magnitude measuring the distance to the decision boundary. The gate g replaces the fixed per-token weight of the logit-mixing rule in Appendix B.6 and is evaluated once after both answers are complete. The routing logit is deterministic rather than calibrated, and no correctness label or judge output enters it. For each source of F_t, the thresholds (τ_D, τ_B, μ) are frozen on development dialogues before evaluation, and Appendix C.4 discusses the complexity and applicability of FSCR.

5 Experiments

To validate our analysis findings and evaluate the proposed FSCR, we conduct comprehensive experiments on explicit-forgetting benchmarks with both annotated and dialogue-derived forgotten states. We employ 4 matched pairs of DeepSeek-R1-Distill models and their base models, namely Qwen-1.5B, Qwen-7B, Qwen-14B, and Llama-8B, as listed in Table A1. All FSCR results are based on the frozen thresholds in Appendix C.3, and no threshold is selected on the held-out evaluation dialogues. Transfer to derived external tasks with 2 additional model families is in Appendix F, and complete control results with clustered intervals are in Appendix D.

5.1 Setting

Benchmarks Original ICF denotes the unmodified ICF-Bench of Section 3.1, in which each of the 994 dialogues ends with a forgetting query after an explicit forget request. ICF-DSM derives 2,982 immediate, delayed, and adversarial forgetting probes from the same source dialogues, while its recall and retained-fact targets serve only the analysis. The full sets include the development dialogues for threshold selection. In contrast, a held-out tail of 434 dialogues with source IDs from 562 onward is excluded from threshold selection and contains 1,302 probes per pair.

Table 2: **Annotated-state FSCR effects.** Gain is the difference in success rate from the reasoning-distilled model, and precision is fixes divided by fixes plus regressions. Candidate pairs whose outcome may change receive 3 fresh votes. Panel (a) gives exact McNemar p-values, and panels (b) and (c) offer dialogue-clustered 95% intervals. The full sets include development dialogues, whereas the tail excludes them.

Pair	Routes (%)	Fix/reg.	Gain (%)	Precision	Inference
(a) Original ICF: 994 final queries per pair					
Qwen-1.5B	50 (5.0)	13/3	+1.01	81.2	$p = 0.0213$
Qwen-7B	64 (6.4)	39/4	+3.52	90.7	$p = 3.1\times10^{-8}$
Qwen-14B	102 (10.3)	93/0	+9.36	100.0	$p = 2.0\times10^{-28}$
Llama-8B	53 (5.3)	31/1	+3.02	96.9	$p = 1.5\times10^{-8}$
(b) ICF-DSM full set: 994 dialogues / 2,982 probes per pair					
Qwen-1.5B	329 (11.0)	48/10	+1.27	82.8	[0.77, 1.78]
Qwen-7B	383 (12.8)	195/9	+6.24	95.6	[5.37, 7.14]
Qwen-14B	414 (13.9)	185/14	+5.73	93.0	[4.80, 6.71]
Llama-8B	256 (8.6)	141/5	+4.56	96.6	[3.82, 5.33]
(c) ICF-DSM held-out tail: 434 dialogues / 1,302 probes per pair					
Qwen-1.5B	145 (11.1)	18/5	+1.00	78.3	[0.31, 1.77]
Qwen-7B	163 (12.5)	78/3	+5.76	96.3	[4.45, 7.07]
Qwen-14B	169 (13.0)	68/7	+4.69	90.7	[3.38, 6.07]
Llama-8B	123 (9.4)	62/2	+4.61	96.9	[3.46, 5.84]

Baselines FSCR keeps y_D by default, and hence the reasoning-distilled model serves as the reference. Both models of each pair decode greedily and independently, and all routing rules share the same candidate answers. We further compare 4 frozen routing controls that change only the routing logit. High leakage only keeps z_{fail} alone, and margin only keeps z_{contrast} alone. Lower leakage routes whenever $L_B < L_D$, while same-coverage random routing matches the coverage of FSCR without leakage evidence.

Evaluation We report the gain in success rate over the reasoning-distilled model, the routing coverage, and the changed-case precision, i.e., fixes divided by fixes plus regressions. Candidate pairs whose outcome may change receive 3 fresh DeepSeek-V4 Pro votes on the visible final answers with a 2-vote majority, and all other answers keep their existing labels. The 3 votes are unanimous on 96.9% to 100% of the judged answers per model, as reported in Table A8. To validate these judgments, 2 annotators blindly audited both answers of 66 routed probes. The annotators reach $\kappa = 0.863$ on forgetting violations, and the automatic SMS matches their SMS on 75.0% to 81.1% of the 132 answers, which supports our behavioral conclusions within the audited sample. Agreement on paired fix and regression labels is lower, whereas both annotators find fewer regressions than the automatic judge, as detailed in Appendix D.5. We apply exact McNemar tests on Original ICF, whereas on ICF-DSM we compute 95% intervals and sign-flip tests that resample source dialogues together with their 3 probes. Computing costs: FSCR adds 1 base-model generation that can run in parallel, and its scoring takes roughly linear time, as analyzed in Appendix C.4.

5.2 Explicit-Forgetting Results

Table 2 presents the explicit-forgetting results, and we can observe that FSCR consistently improves the reasoning-distilled models across benchmarks and model pairs. Concretely, on Original ICF, FSCR routes 5.0% to 10.3% of the final queries to the base model and gains 1.01% to 9.36%, with exact McNemar $p \leq 0.0213$ for every pair. Qwen-14B benefits the most, as all 93 of its changed answers are fixes. On ICF-DSM, FSCR routes 8.6% to 13.9% of the probes and gains 1.27% to 6.24%, and all dialogue-clustered 95% intervals exclude zero. Moreover, fresh judgments on the held-out tail retain gains from 1.00% to 5.76%, with the largest clustered sign-flip p of 0.0110. However, we notice that Qwen-1.5B gains the least and has the lowest changed-case precision, from 78.3% to 82.8%. We attribute this to its base model, which attains the lowest ICF-DSM SMS of 18.65% among the 4 base models as reported in Table A4 and thus offers fewer compliant alternatives. Across 2



(a) Net gain by probe position (b) Logit mixing vs. FSCR (c) Base-only wins

Figure 4: **Where FSCR corrects errors.** Net gains in panel (a) are over the 994 ICF-DSM probes of each position. Panel (b) compares both methods on Qwen-7B, where logit mixing covers 100 probes per position from source IDs 60 to 159 and FSCR covers all 994. In panel (c), light bars count base-only wins on 2 prespecified holdouts of 300 probes, teal bars show FSCR, and orange bars show same-coverage random routing with its central 95% range.

benchmarks and 4 model pairs, FSCR recovers explicit-forgetting failures at a changed-case precision of at least 78.3% without modifying either model, which supports the effectiveness of the earlier analysis. We further apply the frozen thresholds to the MQuAKE-derived knowledge-update task with 2 additional model families in Appendix F. As shown in Figure A3, FSCR improves all 6 model pairs on this task by 0.10% to 4.60%, and 4 of them are significant at unadjusted $p < 0.05$.

5.3 Where FSCR Corrects Errors

To connect these gains with the analysis of Section 3, we further examine where FSCR corrects errors, as illustrated in Figure 4. Figure 4(a) shows that delayed probes contribute the largest share of net corrections in every pair, from 68.4% for Qwen-14B to 100% for Qwen-1.5B. Concretely, FSCR nets 168 delayed corrections for Qwen-7B with only 1 regression, i.e., a gain of 16.90% on its 994 delayed probes. In every pair, delayed corrections also reach the highest changed-case precision, from 95.2% to 100%. This concentration matches finding i, as the advantage of the reasoning-distilled models shrinks or reverses on delayed probes, whereas their base models improve there in every pair. In contrast, immediate probes receive the fewest routes in every pair, from 15 to 101 of 994, which is consistent with the higher immediate SMS of the reasoning-distilled models. Table A7 lists all per-position counts. As Figure 4(b) illustrates, mixing base-model logits introduces 39 regressions per 100 immediate Qwen-7B probes, whereas FSCR introduces only 0.5, which is in line with finding iii. However, immediate corrections remain small and are even slightly negative for Qwen-1.5B, and adversarial corrections are less precise, e.g., 13 fixes against 10 regressions for Qwen-14B.

We then measure how much complementarity FSCR recovers on 2 prespecified holdouts of 300 probes each, which were generated separately from the full sets. As finding ii suggests and Figure 4(c) shows, only the base answer is correct on 24 Llama-8B probes and 9 Qwen-1.5B probes, and FSCR recovers 12 and 5 of them without any regression. In contrast, routing the same number of random answers is expected to lose 1.92 and 2.42 answers, and it reaches the net gain of FSCR with probability at most 3.19×10^{-5}, as reported in Table A8. The remaining base-only wins mark the headroom for routing evidence beyond lexical leakage. These results demonstrate that FSCR corrects errors where the analysis locates them, which also validates our analysis findings.

5.4 Dialogue-Derived Forgotten State

Annotated forgotten state is rarely available in deployment, where the forget scope must be read from the dialogue itself. FSCR exploits a fixed lexical rule, as the routing logit requires the forgotten content rather than its annotation. To validate that, we compare 3 state inputs on the held-out tail in Table 3. All configurations share candidate answers and cached outcome labels, and their estimates are thus not interchangeable with the fresh judgments in Table 2. The annotated and dialogue-derived configurations employ separately frozen thresholds: dialogue-derived state retains gains from 1.38% to 5.45% in every pair, and its pair-averaged gain reaches 84.7% of that with annotated state. Its

Table 3: **Frozen state-input configurations on the 434-dialogue tail.** Each cell gives the SMS gain/changed-case precision, followed by routing coverage C. All rows share candidate answers and cached outcome labels. The annotated and dialogue-derived configurations have separate thresholds.

Configuration	Qwen-1.5B	Qwen-7B	Qwen-14B	Llama-8B
Annotated state	+1.54 / 100.0 C: 11.1	+6.07 / 98.8 C: 12.5	+4.84 / 93.2 C: 13.0	+4.61 / 98.4 C: 9.4
Full source text	+0.69 / 84.6 C: 13.5	+4.07 / 100.0 C: 10.1	+3.76 / 88.9 C: 10.4	+3.53 / 96.0 C: 10.4
Dialogue-derived state	+1.38 / 80.0 C: 25.5	+5.45 / 87.4 C: 23.9	+3.99 / 83.3 C: 24.4	+3.61 / 83.1 C: 15.7

Table 4: **Frozen routing controls on the 434-dialogue tail.** Each cell gives the SMS gain / changed-case precision, with routing coverage in parentheses, all in %. All rows share the cached candidates and labels, and the annotated-state FSCR row repeats Table 3.

Rule	Qwen-1.5B	Qwen-7B	Qwen-14B	Llama-8B
Annotated-state FSCR	+1.54 / 100 (11.1)	+6.07 / 98.8 (12.5)	+4.84 / 93.2 (13.0)	+4.61 / 98.4 (9.4)
High leakage only	+0.61 / 56.9 (70.6)	+5.84 / 76.0 (60.8)	+5.61 / 73.0 (39.6)	+4.61 / 67.4 (61.4)
Lower leakage	+2.38 / 81.6 (41.2)	+8.45 / 91.7 (38.1)	+7.99 / 87.1 (41.2)	+6.68 / 86.0 (28.7)
Margin only	+2.23 / 93.9 (21.4)	+7.76 / 97.2 (20.4)	+6.61 / 92.2 (23.2)	+6.61 / 97.8 (15.2)
Same-coverage random	-0.15 / 42.9 (11.1)	-1.08 / 34.8 (12.5)	-0.23 / 45.9 (13.0)	-0.69 / 39.0 (9.4)

changed-case precision ranges from 80.0% to 87.4%, lower than the 93.2% to 100% with annotated state. Meanwhile, it routes more answers and introduces more regressions in every pair. The full-source-text configuration keeps the annotated scope while replacing annotated units with raw dialogue turns, and its smaller gains suggest that the scorer is sensitive to the textual form of forgotten units. Against a finalized human reference, the extractor attains 72.0% precision, 87.3% recall, and 78.9% F1, while 2 annotators reach $\kappa = 0.829$ on 99 jointly decidable dialogues, as detailed in Appendix E. Its recall exceeds its precision, and hence the extracted state often contains extra dialogue turns, where incidental overlap may make a compliant answer appear leaky and thus trigger additional regressions. These results suggest that FSCR does not rely on annotated forgotten state.

5.5 Precision–Coverage Trade-offs of Routing Margins

We compare FSCR with 4 simpler frozen routing controls on the same held-out tail. Table 4 shows that high leakage only, which ignores the base answer, lowers changed-case precision to between 56.9% and 76.0% while routing 39.6% to 70.6% of answers. Its Qwen-1.5B gain interval even includes zero, as shown in Figure A2. Random routing at the same coverage as FSCR is near zero or harmful. In contrast, the rules that contrast both answers keep changed-case precision of at least 81.6% in every pair. Among them, lower leakage and margin only achieve larger gains from 2.23% to 8.45%, albeit with lower precision and higher coverage than FSCR in every pair. Across the 4 pairs, FSCR reduces observed regressions from 15 to 7, i.e., 53.3% fewer, while retaining 229 of the 317 fixes of margin only, i.e., 72.2%, which prioritizes precision over coverage as counted in Figure A2. This cached-label comparison does not establish superiority over coverage-matched margin only.

6 Conclusion

In this work, we first discover a gap between remembering earlier information and sustaining an updated instruction state in reasoning-distilled models in multi-turn dialogue and explicit-forgetting experiments. We then explore this gap by comparing matched base and reasoning-distilled models, finding that the base models offer complementary answers and corrective signals, while injecting these signals during generation breaks other correct answers. Inspired by that, we propose a novel FSCR strategy to lift the correction from token-level logits to a sequence-level contrastive routing logit based on forgotten-content leakage. Extensive experiments with multiple LLMs on several explicit-forgetting benchmarks illustrate consistent improvements, demonstrating that the proposed FSCR can recover forgetting failures to achieve better forgetting compliance.

AI USE STATEMENT

We employed AI tools to polish the writing of this paper. Part of the data annotation was also assisted by AI, and the resulting annotations were checked by both AI and humans.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. *arXiv preprint arXiv:2308.14508*, 2023. URL https://arxiv.org/abs/2308.14508.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *arXiv preprint arXiv:2305.05176*, 2023. URL https://arxiv.org/abs/2305.05176.

Xinghao Chen, Zhijing Sun, Wenjin Guo, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, and Xiaoyu Shen. Unveiling the Key Factors for Distilling Chain-of-Thought Reasoning. *arXiv preprint arXiv:2502.18001*, 2025. URL https://arxiv.org/abs/2502.18001.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Falcon LLM Team, Iheb Chaabane, Puneesh Khanna, Suhail Mohmad, Slim Frikha, Shi Hu, Abdalgader Abubaker, Reda Alami, Mikhail Lubinets, Mohamed El Amine Seddik, and Hakim Hacid. Falcon-H1R: Pushing the reasoning frontiers with a hybrid model for efficient test-time scaling. *arXiv preprint arXiv:2601.02346*, 2026. URL https://arxiv.org/abs/2601.02346.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. OpenThoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025. URL https://arxiv.org/abs/2506.04178.

Chi Han, Xin Liu, Haodong Wang, Shiyang Li, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Qingyu Yin, Liang Qiu, Changlong Yu, Yifan Gao, Zheng Li, Bing Yin, Jingbo Shang, and Heng Ji. Can language models follow multiple turns of entangled instructions? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 25445–25460, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1387. URL https://aclanthology.org/2025.findings-emnlp.1387/.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024. URL https://arxiv.org/abs/2410.15553.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.830. URL https://aclanthology.org/2023.acl-long.830/.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. *arXiv preprint arXiv:2305.02301*, 2023. URL https://arxiv.org/abs/2305.02301.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792/.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.257. URL https://aclanthology.org/2024.acl-long.257/.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. *arXiv preprint arXiv:2205.11916*, 2022. URL https://arxiv.org/abs/2205.11916.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VKGTGGcwl6.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. Tuning Language Models by Proxy. *arXiv preprint arXiv:2401.08565*, 2024. URL https://arxiv.org/abs/2401.08565.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.151. URL https://aclanthology.org/2023.acl-short.151/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A Task of Fictitious Unlearning for LLMs. *arXiv preprint arXiv:2401.06121*, 2024. URL https://arxiv.org/abs/2401.06121.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to Route LLMs with Preference Data. *arXiv preprint arXiv:2406.18665*, 2024. URL https://arxiv.org/abs/2406.18665.

Yuli Qian, Zechuan Yang, Wenbiao Ding, Hongzhi Li, and Yutao Xie. Do LLMs forget what they should? Evaluating in-context forgetting in large language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hcJywRYc3n.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. *arXiv preprint arXiv:1706.05806*, 2017. URL https://arxiv.org/abs/1706.05806.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine Unlearning Six-Way Evaluation for Language Models. *arXiv preprint arXiv:2407.06460*, 2024. URL https://arxiv.org/abs/2407.06460.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. *arXiv preprint arXiv:2404.13208*, 2024. URL https://arxiv.org/abs/2404.13208.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yaxuan Wang, Chris Yuhao Liu, Quan Liu, Jinlong Pang, Wei Wei, Yujia Bao, and Yang Liu. DRAGON: Guard LLM unlearning in context via negative detection and reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vQLUAkl5SG.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. *arXiv preprint arXiv:2410.10813*, 2024. URL https://arxiv.org/abs/2410.10813.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9eJv5PS27Q.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

The FSCR implementation and offline reproduction artifacts are available at `https://anonymous.4open.science/r/fscr-review-DEF8/`.

A Implementation Details

This appendix provides the details omitted from the main text. Appendix B details the analysis of Section 3, and Appendix C presents the FSCR algorithm together with its scoring details. Appendices D to F report additional FSCR results, the scope audit, and the transfer to external tasks, while Appendices G and H discuss limitations and extended related work. In this section, we describe the models, benchmarks, generation, and evaluation shared by the analysis in Section 3 and the experiments in Section 5.

A.1 Models and Benchmarks

We employ 4 matched pairs of base and reasoning-distilled models, as listed in Table A1. Each reasoning-distilled model is fine-tuned from the base model in the same row, and the 2 models share Transformer block topology, whereas their tokenizers and chat templates remain model-specific. Table A2 summarizes the scale and role of each benchmark.

Table A1: Matched models employed in the analysis and FSCR evaluation. The final 2 columns give the full model names.

Pair	Base model	Reasoning-distilled model
Qwen-1.5B	Qwen2.5-Math-1.5B	DeepSeek-R1-Distill-Qwen-1.5B
Qwen-7B	Qwen2.5-Math-7B	DeepSeek-R1-Distill-Qwen-7B
Qwen-14B	Qwen2.5-14B	DeepSeek-R1-Distill-Qwen-14B
Llama-8B	Meta-Llama-3.1-8B	DeepSeek-R1-Distill-Llama-8B

Table A2: Benchmark scales and roles in the study.

Benchmark	Scale per model	Study role
Original ICF	994 dialogues and final queries	Explicit-forgetting evaluation
ICF-DSM	994 dialogues, 4,970 target prefixes, 2,982 forgetting probes	Dynamic-state diagnosis and FSCR evaluation
MultiChallenge	273 dialogues, including 69 instruction-retention cases	Memory and instruction-retention diagnosis
Modified MTDM	1,079 prefixes, including 240 instruction-retention targets	Additional behavioral diagnosis

Original ICF is the unmodified ICF-Bench (Qian et al., 2026), whose final query serves as the explicit-forgetting endpoint. To separate immediate compliance from compliance after intervening turns, ICF-DSM derives 5 independent target prefixes from each Original ICF source dialogue. The targets test recall before forgetting, immediate suppression, delayed suppression after 2 distractor exchanges, utilization of a retained fact, and adversarial reconstruction. A response to a target is never inserted into the prefix of another target, and hence the targets are evaluated independently.

The MTDM diagnostic is derived from the 273 MultiChallenge dialogues (Deshpande et al., 2025). A model-assisted annotation extracts atomic historical facts, selects substantive user turns, and specifies the facts required at each target. We then generate an answer independently from each prefix ending at a target user turn. The evaluated aggregates contain 1,079 prefixes, including 240 instruction-retention targets. The construction retains a final-turn fallback for dialogues without target annotations. Late instruction consistency is the fraction of instruction-retention answers without a judged contradiction of historical facts among the last $\max(1, \lceil K_i/3 \rceil)$ targets of each source dialogue, where K_i is its number of targets; a single-target dialogue contributes to both early and late segments. This consistency metric does not require all query-relevant facts to be recalled.

For Multi-IF (He et al., 2024), we employ the Chinese and English rows of the released `multiIF_20241018.csv`, selecting solely by the language field. This yields 1,363 items at

steps 1 and 2: 454 Chinese and 909 English. At step 3, 13 English rows have no third-turn prompt or instruction list in the source CSV and are skipped by the benchmark metric, leaving 1,350 items. Results are reported in Table A3.

A.2 Candidate Generation

For all candidate answers, we adopt greedy decoding with temperature 0 and top-$p = 1$, together with the native chat template of each model. The ICF-DSM runner sets a default cap of 16,384 completion tokens, and we lower the cap to 2,048 tokens for the Llama-8B base model. The primary FSCR comparisons reuse the completed answers in the Original ICF and ICF-DSM candidate files, and hence all routing rules and controls share identical candidates. For scoring, we extract the visible final answer as described in Appendix C.1.

A.3 Judging and Statistical Inference

To obtain reliable labels for the answers that routing may change, each potentially changed candidate pair in the primary evaluation receives 3 independent DeepSeek-V4 Pro judgments with thinking disabled, and a 2-vote majority gives the binary outcome. Answers left unchanged by routing keep their existing labels. Thus, absolute totals combine new majority labels on routed examples with existing labels elsewhere, and we report paired gains rather than treat those totals as uniformly rejudged pass rates. We apply exact McNemar tests on Original ICF. On ICF-DSM, confidence intervals and sign-flip tests resample source dialogues and preserve the 3 forgetting probes in each cluster. The full benchmark includes the development dialogues, whereas the common held-out tail of 434 source dialogues with IDs from 562 onward was excluded from threshold selection. We froze candidate answers, code, and routing settings before the fresh judgments on the held-out tail.

B Analysis Details

This section details the definitions and protocols behind the analysis in Section 3, including the state-management outcome, the delay contrast, the representation comparison, the fixed-answer likelihood, and the interventions during generation.

B.1 State-Management Success

To evaluate whether an answer follows the current instruction state rather than the visible history alone, we formally define SMS per dialogue prefix. For a dialogue prefix x_t, let A_t be the active units, I_t the invalidated units, and $F_t \subseteq I_t$ the explicitly forgotten units. Let $A_t^{\mathrm{req}} \subseteq A_t$ contain the active units that the current query requires the answer to employ. For response y, let $U_t(y)$ be the active units employed by the response, $V_t(y)$ the misused invalidated units, and $W_t(y)$ the misused forgotten units. Let $c_t(y)$ indicate consistency without unnecessary refusal. Then SMS is

$$\mathrm{SMS}_t(y) = \mathbf{1}[A_t^{\mathrm{req}} \subseteq U_t(y)]\mathbf{1}[I_t \cap V_t(y) = \varnothing]\mathbf{1}[F_t \cap W_t(y) = \varnothing]c_t(y). \tag{7}$$

The active-unit condition is satisfied when A_t^{req} is empty. In the evaluation metadata, the target-specific `active_units` field records A_t^{req}, rather than all active content in the dialogue. We further define the response-level misuse indicator of forgotten information as $m_t(y) = \mathbf{1}[F_t \cap W_t(y) \neq \varnothing]$. Avoiding misuse is less restrictive than satisfying SMS, since the answer may still omit a required active fact. The late complementary-outcome groups employ the final 2 ICF-DSM targets and m_t, and they therefore represent misuse rather than full-answer success.

B.2 Paired Delay Contrast

To compare how the 2 models change from immediate to delayed probes, we compute a paired contrast over source dialogues. For source dialogue i and model M, let $s_{i,M}^{\mathrm{imm}}$ and $s_{i,M}^{\mathrm{del}}$ denote the immediate and delayed SMS indicators. The mean paired contrast is

$$\Delta_{\mathrm{delay}} = \frac{1}{N}\sum_{i=1}^{N}\left[(s_{i,D}^{\mathrm{del}} - s_{i,D}^{\mathrm{imm}}) - (s_{i,B}^{\mathrm{del}} - s_{i,B}^{\mathrm{imm}})\right]. \tag{8}$$

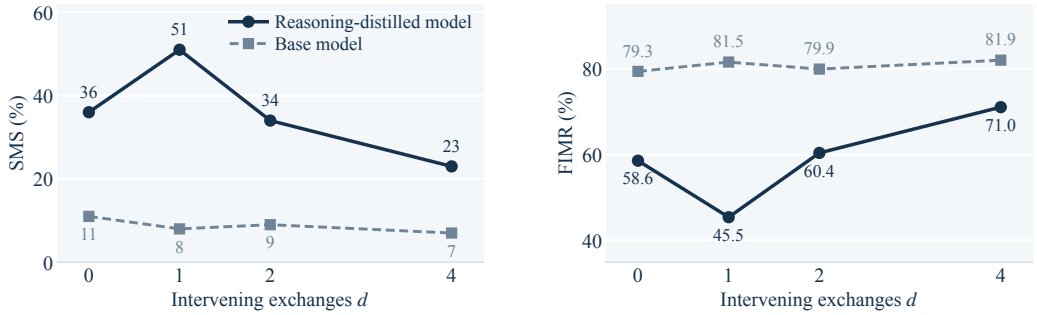
Bootstrap draws resample complete dialogues and retain both models and target positions together. A negative value means that the reasoning-distilled model has a less favorable immediate-to-delayed change than its base model. Since the targets differ in wording and the intervening exchanges are synthetic, this contrast does not isolate dialogue length as a cause.

Beyond explicit forgetting, Table A3 records other task outcomes that qualify a general claim about instruction following, and Table A4 separates global SMS, late-target misuse, and the paired delay contrast. We can observe that overall DSM SMS rises in 3 pairs, while the immediate-to-delayed SMS contrast remains negative in every pair. Note that the late metrics employ retained-fact and adversarial targets, and hence they are not delayed-suppression scores.

B.3 Matched-Query Delay Control

To separate query wording from intervening dialogue, we evaluate the same 100 held-out ICF-DSM source dialogues with 0, 1, 2, and 4 intervening exchanges. Each exchange contains one user message and one assistant reply. Within each source, the pre-delay history, forget request, active and forgotten units, and final query remain identical. Each condition uses the first d exchanges from the same 4-exchange sequence. We report the base and reasoning-distilled models of Qwen-7B, comprising 800 answers across 8 conditions. All sources lie in the held-out tail, with source IDs from 899 to 999 excluding 947. This diagnostic is separate from the original immediate-to-delayed comparison and FSCR threshold selection.

Each answer receives 3 DeepSeek-V4 Pro judgments with thinking disabled and majority aggregation. All 800 reported answers have 3 valid votes, with no missing judgments or unknown outcomes. We report SMS and forgotten-information misuse rate, denoted FIMR. For each response, FIMR is the fraction of annotated forgotten units judged to be misused, averaged across responses. It therefore differs from the binary response-level misuse indicator in Appendix B.1. Active use measures the fraction of active units used. Over-refusal and token-cap rates are reported separately. The completion cap in these runs is 1,024 tokens.



(a) SMS (b) FIMR

Figure A1: **Matched-query delay control on Qwen-7B.** Each point averages 100 answers from the same held-out sources after d intervening exchanges, and all scores are in %. FIMR averages the fraction of forgotten units misused per answer, and hence lower is better. Across the 8 conditions, active use stays between 97% and 100%, over-refusal at or below 1%, and the token-cap rate at or below 5%.

For each source and model, we fit a linear slope to the 4 outcomes at $d \in \{0, 1, 2, 4\}$ and compare the distilled and base slopes. For outcome z, the source-specific slope is

$$b_{i,M} = \frac{\sum_{d\in\{0,1,2,4\}}(d - 1.75)z_{i,M,d}}{8.75}, \qquad \Delta_b = \frac{1}{100}\sum_{i=1}^{100}(b_{i,D} - b_{i,B}). \tag{9}$$

Bootstrap intervals and sign-flip tests operate on complete source dialogues, retaining both models and all 4 conditions together. We also report the paired $d = 2$ minus $d = 0$ contrast. Reported p-values are unadjusted.

Table A3: **Behavioral boundaries beyond explicit forgetting.** Changes are absolute differences in percent-valued scores. Paired step-3 scores list the base model first and the reasoning-distilled model second. In our evaluation of Multi-IF (He et al., 2024), we use 1,363 items at steps 1 and 2 and 1,350 at step 3. Preference and revision retain their original task-specific metrics. MTDM degradation measures the within-dialogue decline and is distinct from the late score. These results qualify the instruction-retention finding rather than establish a common cross-benchmark ranking.

Metric	Qwen-1.5B	Qwen-7B	Qwen-14B	Llama-8B
(a) Multi-IF: multilingual instruction following				
Δ step 1	+7.41	+14.24	-7.13	+19.43
Δ step 2	+15.16	+18.93	+0.81	+21.39
Δ step 3	+12.02	+13.38	-0.19	+15.43
Step 3: paired scores	12.73 / 24.75	13.44 / 26.82	28.94 / 28.75	12.66 / 28.09
(b) Other ICF tasks: reasoning-distilled model minus base model				
Dynamic Preference: retention	-7.79	+6.13	+14.43	+37.04
Subtask Revision: retention	+19.44	+35.49	+33.52	+49.86
Subtask Revision: joint	+21.14	-0.23	+3.01	-14.28
(c) Modified MTDM: additional paired-model contrasts				
IR degradation (higher is worse)	+14.77	+2.79	+2.50	+9.08
Overall late memory	+9.61	+16.14	+13.51	+18.04
Update-retention subtask	-3.33	-13.33	0.00	-10.00

Table A4: **ICF-DSM behavior and temporal contrast.** Panel (a) gives percent-valued scores over the original 994-dialogue, 5-target evaluation. Each paired cell lists the base model first and the reasoning-distilled model second. Late metrics and panel (c) cover retained-fact and adversarial targets only ($n = 1,988$ per pair), excluding delayed suppression. Global FIMR is unit weighted. Late-FIMR is response-level misuse incidence. Panel (b) compares 994 paired immediate and delayed targets per model. DiD and paired 95% intervals are absolute differences in percent-valued scores. The model-only groups in (c) refer to misuse rather than full-answer failure. The final column partitions the reasoning-distilled-model-only group by adversarial and retained-fact target.

(a) Metric: paired scores	Qwen-1.5B	Qwen-7B	Qwen-14B	Llama-8B
SMS	18.65 / 16.12	27.28 / 28.25	49.80 / 50.66	28.29 / 32.21
Late-SMS	7.34 / 6.94	10.51 / 8.90	12.78 / 13.93	11.07 / 9.41
Pre-forget recall	74.79 / 55.24	79.73 / 73.32	82.60 / 80.86	83.05 / 76.12
FIMR	59.17 / 68.61	49.85 / 54.47	23.01 / 26.13	52.55 / 51.93
Late-FIMR	48.14 / 64.29	43.66 / 55.03	32.65 / 38.38	45.67 / 56.49
Retained-fact use	36.62 / 34.54	41.89 / 42.66	43.52 / 47.90	45.76 / 48.43

(b) Pair	Base model SMS, Imm. / Del.	Reasoning-distilled model SMS, Imm. / Del.	SMS DiD [95% CI]	FIMR DiD [95% CI]
Qwen-1.5B	2.01 / 11.17	7.14 / 11.27	-5.03 [-7.95, -2.01]	+5.63 [+2.62, +8.75]
Qwen-7B	5.43 / 39.64	34.00 / 22.43	-45.77 [-50.20, -41.25]	+46.38 [+41.85, +51.01]
Qwen-14B	64.89 / 84.71	76.56 / 74.75	-21.63 [-26.26, -17.00]	+21.53 [+17.10, +26.06]
Llama-8B	3.32 / 40.85	40.44 / 33.20	-44.77 [-49.60, -40.14]	+44.97 [+40.14, +49.60]

(c) Pair	Both clean	Reasoning-distilled model only	Base model only	Both misuse	Adversarial / Retained fact
Qwen-1.5B	632	399	78	879	108/291
Qwen-7B	828	292	66	802	133/159
Qwen-14B	1078	261	147	502	184/77
Llama-8B	736	344	129	779	188/156

Figure A1 plots the Qwen-7B trajectories. From $d = 0$ to $d = 4$, SMS decreases from 36% to 23% for the distilled model, while FIMR increases from 58.58% to 71.02%. This matched-query comparison is limited to Qwen-7B and does not establish the same effect across other model pairs.

The 4-condition analysis additionally identifies significant distill-by-delay interactions for Qwen-7B. Relative to its base model, the SMS slope is 3.86% more negative per exchange, and the FIMR slope is 3.87% more positive, with $p = 0.0094$ and $p = 0.0044$, respectively. Table A5 reports slope estimates and confidence intervals for the Qwen-7B pair across all metrics. Table A6 separately reports the paired $d = 2$ minus $d = 0$ contrasts.

Table A5: **Linear trends across four delay conditions.** Values are percentage points per intervening exchange. Interaction is the distilled slope minus the base slope. Intervals and sign-flip tests use 100 source dialogues per pair.

Pair	Metric	Base	Distill	Interaction [95% CI]	p
Qwen-7B	SMS	-0.83	-4.69	-3.86 $[-6.69, -1.03]$	0.0094
Qwen-7B	FIMR	$+0.50$	$+4.38$	$+3.87$ $[1.38, 6.55]$	0.0044
Qwen-7B	Active use	-0.49	-0.60	-0.11 $[-1.06, 0.77]$	0.9074
Qwen-7B	Over-refusal	0.00	-0.17	-0.17 $[-0.51, 0.00]$	1.0000
Qwen-7B	Token cap	$+0.77$	-0.83	-1.60 $[-2.80, -0.60]$	0.0043

Table A6: **Paired $d = 2$ minus $d = 0$ contrasts.** Values are percentage points. Interaction is the distilled change minus the base change. Intervals and sign-flip tests use 100 source dialogues per pair.

Pair	Metric	Base	Distill	Interaction [95% CI]	p
Qwen-7B	SMS	-2.00	-2.00	0.00 $[-13.00, 13.00]$	1.0000
Qwen-7B	FIMR	$+0.54$	$+1.82$	$+1.28$ $[-10.33, 13.31]$	0.8389
Qwen-7B	Active use	-2.00	0.00	$+2.00$ $[0.00, 5.00]$	0.4999
Qwen-7B	Over-refusal	0.00	0.00	0.00 $[0.00, 0.00]$	1.0000
Qwen-7B	Token cap	$+3.00$	-5.00	-8.00 $[-15.00, -3.00]$	0.0157

B.4 Representation Comparison

To examine whether forgetting failures are localized in representation space, we compare the hidden representations of the 2 models layer by layer. We encode a fixed serialized dialogue prefix independently with each model and take the hidden vector of the final input token at every returned backbone stage. Stacking vectors across examples gives matrices X_ℓ and Y_j for stages ℓ and j. With $H = I - \mathbf{1}\mathbf{1}^\top/n$, $K_\ell = HX_\ell X_\ell^\top H$, and $G_j = HY_jY_j^\top H$, linear centered kernel alignment is

$$\mathrm{CKA}(\ell, j) = \frac{\langle K_\ell, G_j \rangle_F}{\|K_\ell\|_F \|G_j\|_F}. \tag{10}$$

The forgetting group contains cases in which the base-model answer has no misuse while the reasoning-distilled answer misuses forgotten content, and the stable-suppression group contains cases in which neither answer misuses it. These groups concern misuse rather than full SMS. We can observe that the equal-pair CKA averages in the forgetting group decline from 0.959 in the first network quarter to 0.792 in the final quarter. However, the final-stage forgetting-minus-stable differences are $+0.014$, $+0.077$, -0.048, and $+0.035$ across the 4 pairs, and this sign change prevents a common failure-specific localization claim. Note that CKA describes geometry across examples, and it does not measure whether a particular instruction is represented or followed.

B.5 Fixed-Answer Likelihood and Activation Replacement

To test whether the base model carries a corrective signal for a fixed answer, we score the same non-leaking base-model response under the native tokenizer and chat template of each model. The answer likelihood excludes prompt tokens and generated reasoning traces, and the reference is a recorded

model answer rather than a human gold response. For activation replacement, we align the donor and recipient at answer-prediction positions. When native token counts differ, the procedure maps positions by relative progress through the answer, which is not exact character-boundary alignment. We scan attention, MLP, and residual outputs across coarse depth bands and individual layers. The reported recovery is the mean finite per-example ratio of the likelihood gain after replacement to the original likelihood gap between the base and reasoning-distilled models. We exclude denominators with magnitude at most 10^{-8} and do not clip ratios. This statistic measures a change in conditional likelihood, and it does not establish a change in complete-answer compliance.

We further examine stable-suppression cases, on which neither historical answer misuses forgotten content. The gap δ on these cases is -0.402, -1.032, -0.912, and -0.471 for Qwen-1.5B, Qwen-7B, Qwen-14B, and Llama-8B, and all 4 paired intervals lie below zero. Therefore, the likelihood gap does not separate failures from stable cases, which is in line with finding ii in Section 3.2.

B.6 Interventions During Generation

To test whether the corrective signal transfers to free generation, we evaluate 3 interventions that act while the answer is being generated.

Online activation replacement. We select sites on a discovery partition that is disjoint by dialogue, and the confirmation tests 1 residual site and 1 MLP site per pair, each on 15 forgetting cases and 15 stable cases. Both models advance on the same generated tokens, and the base-model activation is recomputed at every decoding position. Generation is greedy with a 16,384-token cap and a 32,768-token context limit. We can observe that online residual replacement changes forgetting-case SMS from 7/60 to 9/60 with 6 fixes and 4 regressions, whereas online MLP replacement changes it from 7/60 to 5/60 with no fixes and 2 regressions. On stable cases, the corresponding changes are from 10/60 to 9/60 and from 10/60 to 11/60. Moreover, the 240 patched responses contain 29 cap events, 25 of which come from the 120 residual-patched responses, while the 120 unpatched controls contain no cap events.

Behavior-selected prompt-only blending. To avoid repeated replacement during decoding, a separate scan selects residual sites by free-generation SMS and forgotten-information misuse rate, i.e., FIMR, on discovery cases. It blends the residual output at the last prompt position once with $\lambda = 0.5$ and then removes the hook. The frozen sites are layer 10 for Qwen-1.5B, layer 6 for Qwen-7B, layers 16 and 44 for Qwen-14B, and layer 6 for Llama-8B, and each site is confirmed on 50 forgetting cases and 50 stable cases. The Qwen-1.5B site raises forgetting-case SMS from 4% to 10% and lowers FIMR from 52% to 30%. However, it also raises the cap rate from 6% to 38% and the mean length from 1,681 to 7,147 tokens, while the other sites change forgetting-case SMS by -6 to $+2$ percentage points.

Base-guided logit mixing. Finally, we mix base-model logits into the answer. The reasoning-distilled model first generates its reasoning until the `</think>` boundary, and the base model then receives only the visible dialogue and the shared answer prefix. At answer position j, the mixed logits are

$$\widetilde{\mathbf{z}}_j = (1-\alpha)\,\mathbf{z}_{D,j} + \alpha\,\Pi\mathbf{z}_{B,j}, \tag{11}$$

where Π maps base-model tokens to the reasoning-distilled vocabulary by token string, and both models consume the selected token. The reported control employs $\alpha = 0.9$ on all 300 forgetting probes from Qwen-7B source dialogues 60 to 159. It gives 3 fixes and 39 regressions on immediate probes, 12 fixes and 12 regressions on delayed probes, and 5 fixes and 5 regressions on adversarial probes. Removing 7 logged abnormal outputs still leaves 19 fixes and 55 regressions.

In summary, these interventions fix some forgetting failures, whereas they also break correct answers or inflate response lengths before the original answer can be inspected, which motivates the sequence-level routing logit of FSCR in Section 4.

C FSCR Algorithm

Algorithm 1 summarizes the complete FSCR procedure of Section 4, and the following subsections detail its leakage scoring, the dialogue-derived forgotten state, its hyperparameters, and its complexity.

Algorithm 1 Forgotten-State Contrastive Routing

Input: dialogue prefix x_t, base model M_B, reasoning-distilled model M_D, annotated forgotten state F_t if available, hyperparameters $\omega_u, \omega_b, \omega_n, \nu, \eta, \rho$, thresholds τ_D, τ_B, μ
Output: routed answer $\widehat{y}_t$

$y_B \leftarrow M_B(x_t),\ y_D \leftarrow M_D(x_t)$ ▷ independent decoding with native token logits
if F_t is not available **then**
  $F_t \leftarrow$ dialogue-derived state extracted from x_t ▷ before either answer is inspected
end if
if $F_t = \varnothing$ **then**
  return y_D
end if
for $y \in \{y_B, y_D\}$ **do**
  $\bar{y} \leftarrow$ suffix of y after the final `</think>` marker, or y if no marker exists
  for $f \in F_t$ **do**
    compute u, b, n by Equation 12 and the exact match e with minimum length ν
    $s(\bar{y}, f) \leftarrow \min\{1, \max[e(\bar{y}, f),\ \omega_u u(\bar{y}, f) + \omega_b b(\bar{y}, f) + \omega_n n(\bar{y}, f)]\}$
  end for
  $L(y, F_t) \leftarrow \max\{m_{\max},\ \rho\, m_{\text{mean}} + (1-\rho)\, m_{\text{count}}\}$ ▷ m_{count} counts units with $s \geq \eta$
end for
$z_{\text{route}} \leftarrow \min\{L_D - \tau_D,\ \tau_B - L_B,\ L_D - L_B - \mu\}$
if $z_{\text{route}} \geq 0$ **then**
  return y_B
else
  return y_D
end if

C.1 Leakage Scoring

Annotated-state FSCR employs ICF-DSM forgotten units, and Original ICF queries obtain the corresponding units through source-dialogue IDs. For answer y and forgotten unit f, let U, B, and N be the sets of distinct content unigrams, adjacent filtered bigrams, and numeric tokens. The coverages in Equation 3 take the forgotten unit as denominator:

$$u(y,f) = \frac{|U(y) \cap U(f)|}{|U(f)|}, \quad b(y,f) = \frac{|B(y) \cap B(f)|}{|B(f)|}, \quad n(y,f) = \frac{|N(y) \cap N(f)|}{|N(f)|}, \tag{12}$$

where empty denominators contribute zero. To normalize text, the lexical scorer lowercases it and extracts alphanumeric tokens. It then applies fixed suffix rules for common plural and verbal endings, removes fixed stopwords, and drops short nonnumeric tokens. The exact-match branch instead keeps unstemmed words and stopwords, joins them with single spaces, and requires a complete normalized unit of at least ν characters to occur contiguously in the answer. This branch is available only when the unit has a retained content token. For both answers, the scorer takes the suffix after the final `</think>` marker when one exists and otherwise keeps the returned text, and an empty forgotten state yields zero leakage and retains the reasoning-distilled answer. Note that this fallback does not detect every malformed reasoning segment, and a low score can reflect a refusal, a non-answer, or a paraphrase as well as correct suppression. Therefore, complete answer quality is judged separately.

C.2 Dialogue-Derived Forgotten State

When no annotation is available, we extract the forgotten state from the visible dialogue before inspecting either answer. The extractor first finds the latest user turn containing an explicit forget verb or a directive that prohibits further usage. It then scores each earlier nonempty turn by the overlap of its alphanumeric words with the scope words in that request, adding 0.03 for assistant turns, and initially retains turns above a fixed relative and absolute overlap threshold. Broad-scope requests expand the selected span through the pre-forget history, and the highest-scoring earlier turn provides a fallback when nothing is selected. The extractor requires an explicit match and does not infer an implicit request. Selected raw turns in chronological order form $\widehat{F}_t$. Therefore, source-text units and annotated units are different textual surfaces even when they refer to the same scope.

C.3 Hyperparameters and Frozen Thresholds

For the lexical scorer, we adopt $\omega_u = 0.55$, $\omega_b = 0.35$, $\omega_n = 0.10$, $\nu = 12$, $\eta = 0.25$, and $\rho = 0.75$ in all experiments. For annotated state, the thresholds are $(\tau_D, \tau_B, \mu) = (0.50, 0.40, 0.15)$. With the extractor fixed, a search on development source IDs 60 to 159 selected the dialogue-derived thresholds $(0.40, 0.50, 0.00)$. The search first preferred coverage from 0.02 to 0.35 and then maximized the number of fixes minus 2 times the number of regressions, with deterministic tie-breaking. Both configurations were frozen before the common 434-dialogue tail. Hence, the state-input comparison in Section 5.4 compares complete configurations with separately chosen thresholds rather than the effect of extraction at a single identical threshold.

C.4 Computing Complexity and Applicability

To compare the computation of FSCR with the interventions in Appendix B.6, let T be the number of generated tokens and C_M the per-token decoding cost of model M. The reasoning-distilled model alone costs $O(T\,C_D)$.

- **FSCR** adds 1 base-model generation of cost $O(T\,C_B)$, which runs independently of and in parallel with the reasoning-distilled model. Leakage scoring only matches normalized tokens in roughly linear time with respect to the lengths of both answers and the forgotten state, which is negligible compared with decoding.
- **Token-level logit mixing** also runs the base model over the answer, whereas the 2 models must decode in lockstep because every selected token is fed to both of them, and the vocabulary mapping Π is applied at every answer position.
- **Online activation replacement** likewise runs both models in lockstep and additionally requires hooks on internal activations at every decoding position.

In summary, FSCR has the same order of cost as running both models once, while it requires neither lockstep decoding nor access to weights, activations, or logits, and hence it also applies to black-box models. FSCR suits assistants that must honor explicit forget or revision requests over long dialogues, such as withdrawn personal details or outdated requirements, when a complementary model is available, and precision matters more than coverage. Implicit state changes and paraphrased disclosure lie beyond its lexical evidence.

D Additional FSCR Results

This section complements Section 5 with per-position counts, judge repeatability, random-routing controls, complete state-input results, and a blind human audit of the automatic judgments.

D.1 Corrections by Probe Position

To locate where FSCR corrects errors, Table A7 decomposes the full-set ICF-DSM corrections of Table 2(b) by probe position. We can observe that delayed probes carry most net corrections with at most 2 regressions per pair, whereas immediate and adversarial probes contribute smaller and less precise corrections, which is in line with Section 5.3.

Table A7: **Full-set ICF-DSM corrections by probe position.** Each cell gives fixes / regressions [routes] over the 994 probes of a position. Delayed share is the fraction of the total net corrections contributed by delayed probes, in %. Figure 4(a) plots the corresponding net gains.

Pair	Immediate	Delayed	Adversarial	Total net	Delayed share
Qwen-1.5B	1 / 4 [27]	40 / 2 [120]	7 / 4 [182]	+38	100.0
Qwen-7B	11 / 5 [33]	169 / 1 [199]	15 / 3 [151]	+186	90.3
Qwen-14B	54 / 3 [101]	118 / 1 [169]	13 / 10 [144]	+171	68.4
Llama-8B	5 / 4 [15]	117 / 0 [137]	19 / 1 [104]	+136	86.0

D.2 JUDGE REPEATABILITY

Since the primary evaluation relies on model judgments, we examine how often the 3 DeepSeek-V4 Pro votes agree on the full-set routed candidates, as reported in panel (a) of Table A8. Unanimity ranges from 96.9% to 100% on Original ICF and from 97.7% to 100% on ICF-DSM. This supports repeatability under the same judge, whereas it does not establish correctness or cross-judge validity. Appendix D.5 therefore compares these judgments with a blind human audit.

Table A8: **Judge repeatability and routing headroom.** Panel (a) measures repeatability within the same judge rather than accuracy or cross-judge robustness. Panel (b) employs 2 prespecified holdouts of 300 probes each. The random range is a central 95% interval of the exact same-coverage distribution, and p_{random} is its probability of matching or exceeding FSCR. Oracle capture counts the recovered base-only wins.

		Unanimous votes	
Benchmark / pair	Pairs judged	Base model	Reasoning-distilled model
(a) 3-vote agreement on full-set routed candidates			
Original ICF / Qwen-1.5B	50	50/50 (100.0%)	50/50 (100.0%)
Original ICF / Qwen-7B	64	62/64 (96.9%)	64/64 (100.0%)
Original ICF / Qwen-14B	102	102/102 (100.0%)	102/102 (100.0%)
Original ICF / Llama-8B	53	53/53 (100.0%)	53/53 (100.0%)
ICF-DSM / Qwen-1.5B	329	326/329 (99.1%)	328/329 (99.7%)
ICF-DSM / Qwen-7B	383	382/383 (99.7%)	380/383 (99.2%)
ICF-DSM / Qwen-14B	414	409/414 (98.8%)	414/414 (100.0%)
ICF-DSM / Llama-8B	256	250/256 (97.7%)	255/256 (99.6%)

Pair [routes]	FSCR net	Random net [95% range]	p_{random}	Oracle capture
(b) Separate prespecified holdouts: 300 probes per pair				
Llama-8B [23]	+12	-1.92 [-6, 2]	3.31×10^{-10}	12/24 (50.0%)
Qwen-1.5B [29]	+5	-2.42 [-6, 1]	3.19×10^{-5}	5/9 (55.6%)

D.3 RANDOM ROUTING AND BASE-ONLY HEADROOM

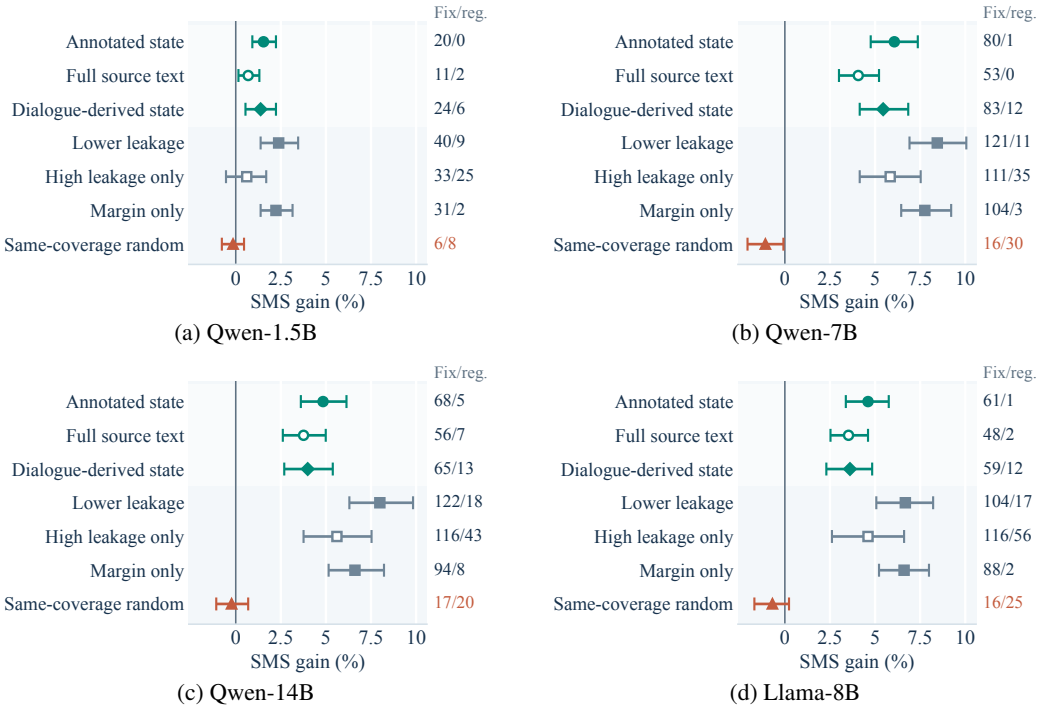
To test whether the gains come from switching itself, panel (b) of Table A8 compares FSCR with same-coverage random routing on 2 holdouts of 300 probes each. The holdouts are drawn from source IDs 362 to 461 for Llama-8B and 462 to 561 for Qwen-1.5B. Both holdouts were prespecified and generated separately from the full-set candidates, and the routing rule was frozen before any judgment. The random distribution is the exact distribution of net corrections when the same number of answers is routed uniformly at random. The oracle counts base-only wins, i.e., probes on which only the base answer is correct, and it is not deployable because it requires correctness labels. We can observe that FSCR recovers about half of the base-only wins without any regression, whereas random routing is expected to reduce net corrections.

D.4 COMPLETE STATE-INPUT AND CONTROL RESULTS

Figure A2 plots the gains with dialogue-clustered intervals and lists the fixes and regressions for all configurations in Tables 3 and 4. These configurations share cached labels, and hence their annotated-state estimate should not be substituted for the fresh judgments in Table 2(c). Moreover, the intervals show that high leakage only on Qwen-1.5B is the only nonrandom configuration whose gain interval includes zero.

Interpreting the frozen-gate trade-off. All counts in this comparison use the shared cached labels in Figure A2, rather than the fresh judgments in Table 2(c). The fixed FSCR rule retains the same contrast threshold $\mu = 0.15$ as margin-only and adds two absolute leakage checks, so its routed set is a subset of that rule's routed set. The pooled counts are 229 fixes and 7 regressions for FSCR, versus 317 and 15 for margin-only. Among the excluded routes whose SMS outcome

changes, the additional checks remove 88 fixes and 8 regressions, corresponding to changed-case precision of 91.7%, compared with 97.0% for the retained FSCR routes. This quantifies the cost of more selective routing: 53.3% fewer observed regressions while retaining 72.2% of margin-only's fixes, with net corrections decreasing from 302 to 222. These summaries pool 5,208 model–probe evaluations sharing 434 source dialogues. They are descriptive arithmetic comparisons, not additional independent experiments or tests of significance. They neither establish dominance over a coverage-matched margin rule nor guarantee deployment risk.



(a) Qwen-1.5B (b) Qwen-7B

(c) Qwen-14B (d) Llama-8B

Figure A2: **Complete cached-label state-input and routing-control results.** Each panel covers the 1,302 probes of the 434-dialogue tail for 1 pair. Markers give the SMS gain over the reasoning-distilled model and whiskers give dialogue-clustered 95% intervals, both as absolute differences in %. Teal markers denote FSCR with the 3 state inputs of Table 3, gray markers denote the leakage-based controls of Table 4, and orange markers denote same-coverage random routing. The right column lists fixes and regressions, and both tables report routing coverage and changed-case precision.

D.5 Blind Human Audit of Automatic Judgments

To validate the automatic SMS judgments rather than the forget-scope extraction, we conduct an independent blind audit with 2 annotators on ICF-DSM probes routed by the frozen FSCR gate. The audit samples 66 comparison groups with fixed quotas stratified by model pair, probe position, and automatic transition. Each transition from the reasoning-distilled answer to the base answer is a fix, a regression, an unchanged failure, or an unchanged success. The groups cover all 4 pairs with 23 immediate, 20 delayed, and 23 adversarial probes. Among them, 60 groups come from the held-out tail, whereas the remaining 6 are recorded fallbacks from outside the tail for strata with too few eligible tail sources.

Each group presents the complete dialogue context, the current query, and the 2 answers as anonymous candidates, whose order is randomized independently for each annotator. The annotators see no model identity, automatic label, transition, leakage score, or source ID, and they label every candidate on its own without consulting each other. For each candidate, they judge 5 core fields, namely answer correctness, forgetting violation, whether the active requirement is met, over-refusal, and contradiction, and they further record the main error category and a confidence score from 1 to 3. The human SMS of a candidate passes only when the answer is correct, meets the active requirement,

and shows no forgetting violation, over-refusal, or contradiction. We report the raw agreement and Cohen's κ of each field, and we compare the human SMS with the automatic SMS from the 3-vote majority. Since the quotas are fixed within strata, the audit characterizes the reliability of the judgments rather than the natural rates of the benchmark.

Table A9: **Blind human audit of the automatic judgments.** Panel (a) compares the 2 annotators on 132 candidates, where κ is undefined when both annotators assign a single class. Its last column counts yes labels for the 5 core fields, forgotten leakage for the error category, and the highest confidence of 3. Panel (b) compares the automatic 3-vote SMS with the composite SMS of each annotator. Panel (c) counts paired transitions from the reasoning-distilled answer to the base answer over 66 comparison groups, and the consensus row excludes the 15 groups on which the annotators disagree. The audit employs stratified quotas, and hence these counts do not estimate natural rates.

(a) Field	Agreement (%)	Cohen's κ	Counted labels, A / B
Answer correct	87.88	0.740	55 / 39
Forgetting violation	93.18	0.863	69 / 72
Active requirement met	97.73	0.393	131 / 128
Over-refusal	100.00	n/a	0 / 0
Contradiction	93.18	0.000	0 / 9
Error category	84.85	0.743	69 / 72
Confidence	83.33	0.538	93 / 111

(b) Annotator	Human fail auto fail	Human fail auto pass	Human pass auto fail	Human pass auto pass	Agreement (%)
A	65	12	13	42	81.06
B	69	24	9	30	75.00

(c) Transitions	Fix	Regression	Unchanged fail	Unchanged pass	Agreement with automatic (%)
Automatic 3-vote	23	19	18	6	n/a
Annotator A	27	10	20	9	66.67
Annotator B	22	7	32	5	57.58
Consensus	20	6	20	5	n/a

Table A9 reports the results, and we can observe that neither annotator marks any core field as uncertain or assigns the lowest confidence, while the highest confidence of 3 covers 93 and 111 of the 132 candidates. The annotators agree on forgetting violations for 93.18% of the candidates with $\kappa = 0.863$, while answer correctness and the error category reach κ of 0.740 and 0.743. Forgotten leakage is the most frequent error category for both annotators, covering 69 and 72 of the 132 candidates. In contrast, almost every candidate meets the active requirement and shows no over-refusal or contradiction, hence, these imbalanced fields have low or undefined κ despite raw agreement of at least 93.18%. The automatic SMS agrees with the human SMS on 81.06% and 75.00% of the candidates for the 2 annotators. The automatic judge passes 54 candidates, which is close to the 55 of annotator A, whereas annotator B is stricter and passes only 39.

However, we notice that the paired transitions agree with the automatic ones on only 66.67% and 57.58% of the groups, since each transition compounds the judgments of 2 candidates. The 2 annotators reach the same transition on 51 of the 66 groups, and the stricter annotator B labels 32 unchanged failures against 20 for annotator A. Moreover, both annotators label fewer regressions than the automatic judge, 10 and 7 against 19, while they label 27 and 22 fixes against 23. Hence, within the audited sample, the automatic labels do not appear to overstate the net corrections of FSCR.

In summary, the blind audit supports the automatic judgments of forgetting violations and, within the audited sample, the behavioral conclusions of FSCR. However, individual fix and regression labels carry nonnegligible judging noise, and the stratified counts should not be read as natural error rates.

E SCOPE ANNOTATION AND AUDIT

To assess the reference for dialogue-derived state, 2 annotators selected pre-forget interaction anchors on the same 100 dialogues and separately judged whether each scope was decidable. The finalized reference excludes 1 dialogue that was not jointly decidable. Exact anchor agreement differs from semantic scope agreement, because benchmark metadata can identify the user turn that elicited a fact while the extractor selects the assistant response containing it. Therefore, Table A10 reports anchor agreement, a separate model-assisted semantic audit, and accuracy against the finalized human reference without pooling these evaluations. We can observe that the 2 annotators agree on 506 of 553 anchor decisions with $\kappa = 0.829$, whereas exact scope-set agreement is 72.7%, since a single differing anchor breaks an exact match.

Table A10: **Scope-reference agreement and extraction accuracy.** Panel (a) compares the 2 human annotators. Panel (b) evaluates the extractor against 3 distinct references. These protocols assess different selections and must not be pooled.

Scope-set measure	Value	Anchor measure	Value
(a) Two-annotator agreement			
Annotated dialogues	100	Eligible decisions	553
Jointly decidable dialogues	99	Both include / exclude	273 / 233
Exact scope-set agreement	72/99 (72.7%)	A-only / B-only	25 / 22
Mean dialogue-level Jaccard	0.890	Anchor-level agreement	91.5%
Set-membership micro-F1	92.1%	Cohen's κ	0.829

Reference / evaluation set	Precision	Recall	F1
(b) Extractor accuracy (%), reference-specific evaluation			
Benchmark anchors / 434-dialogue tail	26.3	46.3	33.5
Model-assisted semantic audit	64.2	95.2	76.7
Finalized human reference	72.0	87.3	78.9

F TRANSFER TO EXTERNAL FORGETTING TASKS

To examine whether FSCR transfers beyond ICF, we apply the frozen annotated-state thresholds to explicit-forgetting tasks derived from the MQuAKE-derived knowledge-update task, together with 2 additional model families.

F.1 TASK CONSTRUCTION

MQuAKE-derived dialogues employ 1,000 unique knowledge cases with 350 2-hop, 350 3-hop, and 300 4-hop chains (Zhong et al., 2023). Each dialogue presents an original chain, invalidates old information, and supplies replacement facts. This construction evaluates in-context knowledge updating rather than parametric knowledge editing. The task contains 500 immediate, 300 delayed, and 200 adversarial probes, with exactly 1 target per source. The released packet designates 700, 150, and 150 cases as training, development, and test sets, respectively. Following an initial 150-case pilot, we evaluate the frozen rule on all 1,000 cases.

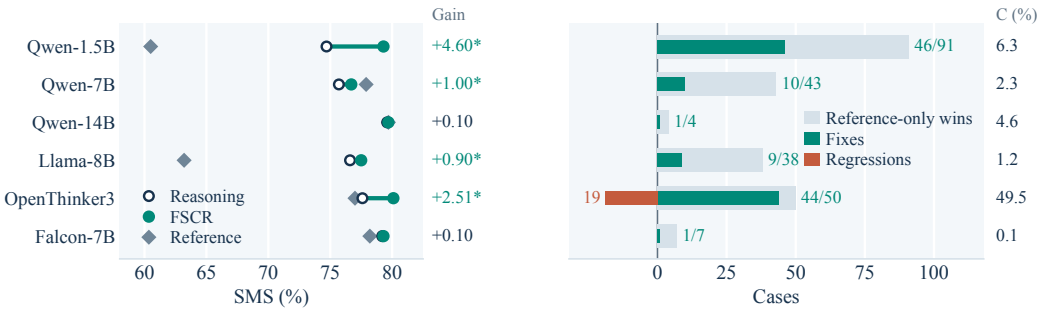
F.2 FROZEN RULE AND JUDGMENT

The 4 diagnostic pairs and the additional OpenThinker3 and Falcon pairs generate complete candidate answers with native chat templates and greedy decoding. For OpenThinker3, the reference checkpoint is `Qwen/Qwen2.5-7B-Instruct` and the reasoning checkpoint is `open-thoughts/OpenThinker3-7B`. The reference is instruction-tuned, and the reasoning model is further fine-tuned on OpenThoughts3 (Guha et al., 2025). For Falcon, the reference checkpoint is `tiiuae/Falcon-H1-7B-Base` and the reasoning checkpoint is `tiiuae/Falcon-H1R-7B`, a public model trained with reasoning SFT and RL (Falcon LLM

Team et al., 2026). Their public checkpoint sources are the corresponding model repositories.[1] The available execution scripts record local checkpoint paths rather than immutable checkpoint revisions. These additional pairs are transfer tests rather than controlled comparisons of reasoning distillation. The same annotated-state thresholds $(0.50, 0.40, 0.15)$ score full forgotten-unit descriptions. Since the experiment record does not establish the 3-vote judging protocol of the primary ICF evaluation, external SMS relies on separately cached final-answer labels. We intersect candidate and label IDs, leaving 998 to 1,000 aligned MQuAKE cases per pair. Each source contributes once, and we apply exact paired McNemar tests. The reported p values are unadjusted for multiple model-pair comparisons.

F.3 Results on MQuAKE

Figure A3 reports SMS scores, routing coverage, fixes and regressions, and available headroom on the MQuAKE-derived task. FSCR improves SMS across all 6 model pairs using the same frozen thresholds, with gains ranging from 0.10 to 4.60 percentage points. Four pairs achieve statistical significance at unadjusted $p < 0.05$. Qwen-1.5B achieves the largest gain with 46 fixes and no regressions. OpenThinker3 gains 2.51 percentage points with 44 fixes and 19 regressions at 49.5% routing coverage. Qwen-14B and Falcon each gain 1 net correction. These results demonstrate transfer of the frozen routing rule beyond ICF to knowledge-update dialogues.



(a) SMS per model pair (b) Recovered reference-only wins

Figure A3: **Frozen-rule FSCR transfer to the MQuAKE-derived task.** Panel (a) shows the SMS of the reference model, the reasoning model, and FSCR on 998 to 1,000 aligned cases per pair, and its right column gives the gain in %, where asterisks mark unadjusted exact McNemar $p < 0.05$. In panel (b), light bars count cases where only the reference answer is correct, teal and orange bars count the fixes and regressions of FSCR, and the right column gives routing coverage C. All pairs employ the frozen annotated-state thresholds $(0.50, 0.40, 0.15)$ and complete forgotten-unit descriptions.

G Limitations

Although FSCR has been demonstrated to be effective on explicit-forgetting benchmarks, it requires a recognizable forget request and a second generation. Its lexical evidence can miss paraphrased disclosure, penalize harmless mentions, or favor a refusal that omits an active requirement, and hence low leakage does not prove compliance. Synthetic intervening turns, derived external annotations, and model-based judgments further limit interpretation, while transfer to derived external tasks depends on the model family, as shown in Appendix F. A blind human audit supports the model-based judgments, whereas it shows lower agreement on paired fix and regression labels, as reported in Appendix D.5. Our experiments cover models from 1.5B to 14B parameters, lacking validation on larger models. Additionally, although each reasoning-distilled model is fine-tuned from its base model, the distillation data also contain non-reasoning samples (DeepSeek-AI, 2025), while the chat templates and some model configurations differ. Hence, the behavioral contrast does not isolate the causal effect of reasoning distillation alone.

[1] Public model cards: Qwen reference, OpenThinker3, Falcon reference, and Falcon-H1R.

H Extended Related Work

Instruction-following benchmarks evaluate constraint satisfaction and multi-turn consistency (Zhou et al., 2023; Deshpande et al., 2025; Han et al., 2025; Laban et al., 2026). Instruction-hierarchy training teaches models to prioritize privileged instructions (Wallace et al., 2024). Long-context and conversational-memory benchmarks emphasize retrieval or updates (Bai et al., 2023; Maharana et al., 2024; Wu et al., 2024). ICF-Bench directly tests in-context forgetting (Qian et al., 2026), which concerns permission to reuse visible content rather than removing training-data influence from weights (Maini et al., 2024; Shi et al., 2024). Representation comparison and activation replacement can characterize model differences (Raghu et al., 2017; Kornblith et al., 2019; Zhang & Nanda, 2024), while candidate ranking and routing offer ways to choose among complete outputs (Chen et al., 2023; Ong et al., 2024; Jiang et al., 2023). In contrast, FSCR derives a sequence-level routing logit from lexical evidence of the forgotten state. Its measured conclusions are limited to the explicit-forgetting tasks and matched models studied in this paper.