Machine Unlearning for Large Language Models: A Survey of Benchmarks, Methods, and Open Problems

Alireza Aalaie *alireza.aalaie11@sharif.edu*
Department of Computer Engineering
Sharif University of Technology

Parsa Sadeghi *sadeghiparsa04@gmail.com*
Department of Mathematical Sciences *parsa.sadeghi27@sharif.edu*
Sharif University of Technology

Danial Parnian *danielparnian@gmail.com*
Department of Computer Engineering
Sharif University of Technology

Omid Ghahroodi *Oghahroodi98@gmail.com*
Department of Computer Engineering *omid.ghahroodi98@sharif.edu*
Sharif University of Technology

Nazanin Yousefi *nazanin.yousefie@gmail.com*
Department of Mathematical Sciences
Sharif University of Technology

Abstract

Machine unlearning for large language models is the problem of removing the influence of designated training data from a trained model without retraining it. This survey reviews papers published between 2023 and 2026, organized as a dependency chain from measurement through execution to verification: the frameworks that define the target, four method paradigms that pursue it, and the studies that test whether the target was reached. The paradigms are output-space objectives, representation-space interventions, weight-preserving methods and robustness-by-design algorithms. What the chain exposes is a gap between what is measured and what is claimed. No method satisfies every evaluation criterion at once, and forgetting is ordered by difficulty, so surface-level success does not imply knowledge-level, privacy-level or robustness-level forgetting. The instruments are fragmented across benchmarks and unstable within them, and knowledge certified as forgotten is recoverable through seven independent channels, from sampling the model more than once to comparing a released model against its predecessor. Each channel can falsify a removal claim and none can verify one, so what these methods demonstrably achieve is suppression of expression rather than removal from the weights. Seven open problems follow, and we propose that verification become a comparison against a declared adversary rather than a certificate.

1 Introduction

Large language models (Brown et al., 2020) are trained on internet-scale corpora and absorb what those corpora contain, including private information, copyrighted text, hazardous technical knowledge and confident falsehoods. *Machine unlearning* is the requirement that follows: that a trained model behave as if designated training data had never been observed (Cao & Yang, 2015; Bourtoule et al., 2021). Four pressures create it. Privacy regulation, most directly the European Union's General Data Protection Regulation, codifies a right

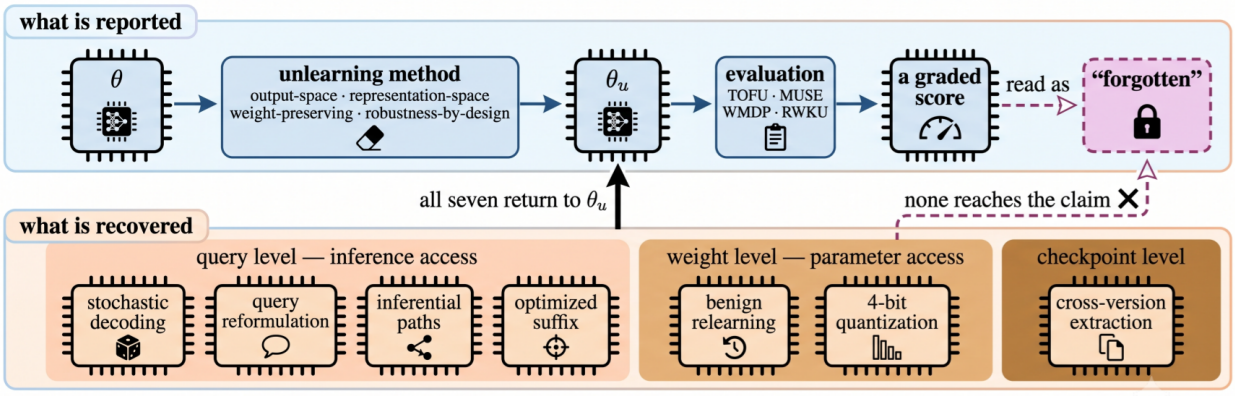

Figure 1: The position this survey argues, drawn. One arrow runs forward and ends in a claim; seven run back into the unlearned model and end nowhere. We read that asymmetry as the central fact about verification in this field: each channel can falsify a removal claim and none can verify one, and the dashed step is the only inference in the diagram that the evidence does not license.

to erasure that can oblige a provider to remove one individual's influence on request (European Parliament and Council of the European Union, 2016). Copyright holders have shown that models reproduce protected passages verbatim (Carlini et al., 2021). The biosecurity and cybersecurity communities have identified models that carry operational knowledge about weapons and software exploitation (Li et al., 2024). And a model that states a falsehood confidently is a quality problem its trainer would like to fix without retraining (Yao et al., 2024).

The requirement is not abstract; it arrives in four concrete forms, each with a different forget set and a different test of success. **Privacy and data removal** takes an individual's memorized information as the target, and because its driver is a statutory right, its test is statistical indistinguishability from a model retrained without that individual (Maini et al., 2024; Shi et al., 2024). **Copyright compliance** targets verbatim reproduction while leaving conceptual understanding of the same domain intact, because the claim is on the expression and not on the subject (Eldan & Russinovich, 2023; Shi et al., 2024). **Safety** targets operational knowledge, and its driver is an adversary rather than a claimant, so the test is that the knowledge cannot be elicited under adversarial conditions (Li et al., 2024). **Factual correction** targets a specific false belief and is driven by output quality rather than by anyone's rights (Yao et al., 2024). What the four share is their structure: a set to forget, a set to keep, and a claim afterwards that the first one is gone.

But meeting any of them in an LLM is structurally harder than the classical case. Retraining is the definition of success and it is unavailable: pre-training runs to trillions of tokens over thousands of GPU-hours and does not decompose into shards that can be dropped and recomputed. What must be removed is also not stored where it can be found: knowledge is distributed across shared parameters rather than localized, so removing one fact degrades unrelated ones. That is the *knowledge entanglement* that makes the forget–retain trade-off the field's recurring difficulty (Maini et al., 2024). Refusal is the tempting shortcut, and a different thing. A system prompt, an RLHF policy (Ouyang et al., 2022) or refusal training changes what a model says while leaving what it knows in the weights, where adversarial prompting or a later fine-tune reaches it (Li et al., 2024). That distinction, between suppressing an expression and removing the information, is what this survey turns on; Section 7 assembles the evidence, and Figure 1 states the position in one picture.

Which is why the literature organizes around three linked questions rather than one: what to measure, how to intervene, and whether the intervention holds. It is also why the answers keep returning the same two findings: no method satisfies every criterion at once, and what the instruments certify is narrower than what the certificates are read to mean. Section 2 fixes the vocabulary and adds two desiderata, Section 3 places the survey against the two prior ones, and Section 4 sets out the chain. Sections 5–7 take the three questions in turn, Section 8 states seven open problems and a proposal, and Section 9 concludes.

This survey contributes:

- a review of the 2023–2026 corpus, organized as a dependency chain in which the measurement literature sets what a method may claim and the persistence results test it
- a formal statement of two desiderata no reviewed paper states, stability and persistence, and seven recovery channels that show persistence violated, ordered by the access an adversary needs: each can falsify a removal claim and none can verify one
- a proposal that verification be reported comparatively: how much of the target a method hides, at what cost, against a named adversary whose access is declared.

1.1 Scope and Paper Selection

The papers answering them were selected on three criteria. The primary contribution had to address LLM unlearning directly: a method, an evaluation framework, or a test of whether removal holds. It had to appear between 2023 and 2026, the window in which this literature separated from classical machine unlearning, which enters only as background (Section 2.1). And it had to engage at least one desideratum of Section 2.3, so its claims sit in a common frame. Thirty papers qualify: five evaluation frameworks and platforms, twenty-one method and analysis papers, and four persistence studies; Figure 2 is the full roster.

2 Background

This section fixes only the vocabulary the review sections use: what unlearning must achieve, how success is stated formally, and which desiderata they measure against.

2.1 Machine Unlearning for Large Language Models

Beginning with why the obvious definition is unavailable. *Exact* unlearning (retraining without the offending data) is available in the classical setting through data partitioning (Bourtoule et al., 2021), which Liu et al. (2025b) and Li et al. (2025) treat at length. LLM pre-training does not decompose: it is one pass over trillions of tokens under a data-mixing schedule, and the parameters admit no attribution back to the examples responsible for them.

Multi-level retention. Infeasibility is not the only problem: what must be removed is itself layered. A model retains verbatim sequences, the knowledge they express, and the capabilities acquired from them, each a stronger requirement than the last: stopping reproduction is not removing the fact, and neither prevents statistical detection of prior exposure (Shi et al., 2024). System prompts, refusal training and RLHF reach only the shallowest: expression is suppressed, the knowledge stays in the weights, and fine-tuning recovers it (Li et al., 2024).

2.2 Problem Formulation

So success has to be defined as approximation rather than equivalence. Write θ for a model trained on $\mathcal{D}$, $\mathcal{D}_f$ for the forget set a request names, $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ for its complement, and $P_\theta(y \mid x)$ for the distribution it assigns to response y given input x. An unlearning algorithm returns θ_u, measured against the *retrained oracle* θ_r trained on $\mathcal{D}_r$ alone. The ideal is indistinguishability:

$$P_{\theta_u}(\cdot \mid x) \;\approx\; P_{\theta_r}(\cdot \mid x) \quad \text{for all inputs } x. \tag{1}$$

No method attains it exactly (Maini et al., 2024; Shi et al., 2024), and it presupposes an adversary who sees only θ_u: given θ too, the divergence leaks $\mathcal{D}_f$ even under exact unlearning (Wu et al., 2025).

Two strengthenings sharpen what *approximately* ranges over. Liao et al. (2026) require the objective to hold not on the listed instances but on $[x]_T$, the equivalence class of queries expressing the same knowledge as x under a family T of semantics-preserving transformations, and so on their union over $\mathcal{D}_f$. Li et al. (2026) strengthen the other coordinate, from the target response to the model's own high-confidence set $\mathcal{B}_\theta(x)$: top-k candidates locally, sampled high-confidence generations globally. The first buys forgetting that does not attach to a surface form; the second, forgetting that mass cannot flow around.

2.3 Desiderata

The four conventional requirements. Which turns the definition into testable desiderata. Four are conventional and taken as given here: *forgetting efficacy* on $\mathcal{D}_f$, *utility preservation* on $\mathcal{D}_r$ and unrelated tasks, *privacy*, in that a membership-inference attack (Shokri et al., 2017) should do no better than chance, and *efficiency* against the cost of retraining. Section 5.2 gives the instruments for each.

Stability. But those four say nothing about what the optimization does to the model, and the field's own methods failed on exactly that. This survey states the missing requirement as a bound on how far unlearning may move the model off the forget set:

$$D_{\mathrm{KL}}\big(P_{\theta_u}(\cdot \mid x) \,\|\, P_{\theta}(\cdot \mid x)\big) \;\leq\; \epsilon \quad \text{for } x \notin \mathcal{D}_f \tag{2}$$

Unconstrained gradient ascent violates it, in the failure named catastrophic collapse: coherence lost on every input (Zhang et al., 2024).

Persistence. And stability says nothing about what happens after deployment. The survey's second addition asks that forgetting survive what is done to a model once it ships. Let $\mathcal{P}$ be a post-deployment procedure: one that transforms the parameters, as quantization and fine-tuning do, or only changes how the distribution is read, as sampling does. Write $\mathcal{P}[\theta_u]$ for the distribution then observed:

$$\mathcal{P}[\theta_u](\cdot \mid x) \;\approx\; P_{\theta_r}(\cdot \mid x) \quad \text{for every admissible } \mathcal{P}. \tag{3}$$

Section 7 shows it violated on seven channels.

2.4 The Forget–Retain Trade-off

These desiderata are not symmetric, and the difference shapes the method families. Ji et al. (2024) say that "things the target LLM should remember are far more intractable than those it needs to forget": the forget objective is defined on a finite, fully specified set, the retain on one that can never be representative. We call the gap the trade-off's *asymmetry*. It motivates getting forgetting from an auxiliary model with those objectives reversed, with no retention term at all (Section 6.3).

3 Related Work

Two recent surveys are close enough in scope to serve as direct comparisons, and each is taken here on its own terms. Liu et al. (2025b) read the field as much as a position paper as a review, across four axes (conceptual formulation, methods, assessment, and applications), and press hardest on the ones they judge neglected: the precise specification of unlearning scope, the joint analysis of data and model influence, and the case for adversarial rather than benign efficacy evaluation. They name weight quantization among the threats to authenticity, cast robust unlearning as a two-player game between a defender and an attacker, and call for *certified unlearning*, sketching a stakeholder-indexed reporting card as the vehicle for it. Their coverage closes at the end of 2024, before the methods that would realize the two-player proposal existed, and before the 2026 work reviewed here.

A second survey organizes the same material differently and closes later. Li et al. (2025) give a systematic taxonomy (direct fine-tuning, localized parameter modification, auxiliary-model approaches, and input- or output-based interventions) paired with a structured account of evaluation measures and benchmarks. Its distinctive branch is adversarial evaluation, and it reaches further into the territory of Section 7 than a claim to differentiation can afford to leave unsaid: the model-intervention subtree already collects relearning through fine-tuning and quantization as tests that separate genuine forgetting from suppression, citing the same two papers this survey cites for two of its seven recovery channels, alongside pruning and activation intervention. It also treats multimodal unlearning at length, which falls outside the scope taken here. Its coverage extends to May 2025.

The overlap between them is substantial and worth stating plainly. All three review TOFU and WMDP; MUSE appears here and in Liu et al. (2025b), RWKU here and in Li et al. (2025). All three cover gradient

ascent, negative preference optimization, representation-level interventions and inference-time alternatives, and the forget–retain trade-off, utility preservation, efficiency and the distinction between suppressing and removing knowledge appear in each. Both also teach the fundamentals, and this survey does not: a reader wanting the classical formulation, the method families in full, or the multimodal literature should read them, and what follows assumes that background rather than rebuilding it. Three openings remain. Neither survey orders the recovery results by the access an adversary needs, neither states the asymmetry that each such result can falsify a removal claim while none can verify one, and neither treats extraction after *exact* unlearning. This survey takes the first as an organizing axis and builds the other two on it. It formalizes two desiderata no reviewed paper states, stability and persistence, and takes the four conventional ones from the surveys above. Six of the reviewed papers appeared in 2026, after both surveys closed, and they fall across all four families this survey organizes by: three defined by where the intervention lands, one by what the result has to survive.

4 Survey Organization

The chain runs from measurement through execution to verification. The benchmarks of Section 5 define the target, the methods of Section 6 pursue it, and the recovery results of Section 7 test whether it was reached, each depending on the one before it. One dependency runs the other way: the methods of Section 6.4 are built against the recovery channels Section 7 documents, which is why verification now precedes part of the execution it tests. Figure 2 maps every reviewed paper onto the chain, several of them onto more than one link. Measurement comes first because the target has to exist before a method can pursue it.

5 Evaluation Frameworks

Four benchmarks define what "working" means, and they define it four different ways. TOFU (Maini et al., 2024) asks whether the unlearned model is distributionally indistinguishable from a retrained oracle, and MUSE (Shi et al., 2024) decomposes the question into six independent desiderata. WMDP (Li et al., 2024) asks whether operational capability in a hazardous domain survives, and RWKU (Jin et al., 2024) asks whether biographical knowledge survives reformulated and adversarial queries. A fifth effort, OpenUnlearning (Kurmanji et al., 2025), contributes no corpus and standardizes the running of the others, and method papers add narrower datasets to isolate one mechanism (Appendix A.4). The four definitions are not equivalent, and some method papers define success scenario by scenario instead (Yao et al., 2024): a method can satisfy one definition while failing another. So the designs come first (Section 5.1), and the verdicts after.

5.1 Benchmark Design and Data Regimes

The paradigms differ first in where the forgotten knowledge came from. TOFU and MUSE share a regime: in both, the knowledge to be removed was put there by fine-tuning on a controlled dataset. Inside it they trade experimental control against ecological validity in opposite directions. Maini et al. (2024) buy control. TOFU's corpus is entirely synthetic: fictitious author profiles, each described by generated question–answer pairs, on which base models are fine-tuned until the fictitious facts are reliably memorized (sizes in Appendix A.5). Because the authors do not exist, provenance is exact: if the model knows a fact about one of them, the fine-tuning data is the only place it can have come from. That buys a clean retrained oracle, a scaling axis in forget sets of 1%, 5% and 10% of the authors, and four evaluation subsets that separate the forget set from the retain set, from questions about real authors, and from general world facts. What it costs is realism: fictitious-author QA is not a medical record, a source file or a copyrighted novel. Shi et al. (2024) buy validity instead. MUSE grounds evaluation in real text (a *NEWS* corpus of articles with disjoint articles as the retain set, and a *BOOKS* corpus using the Harry Potter novels against Harry Potter FanWiki), and keeps the retain data used for regularization disjoint from the retain data used for evaluation, so that no method can overfit its own test. Real text is what deletion requests are made of; the cost is that provenance can now only be inferred, and the evidence is two 7B models on two corpora, a limit MUSE states itself. What neither end of the trade buys is a verdict about knowledge the model was not given on

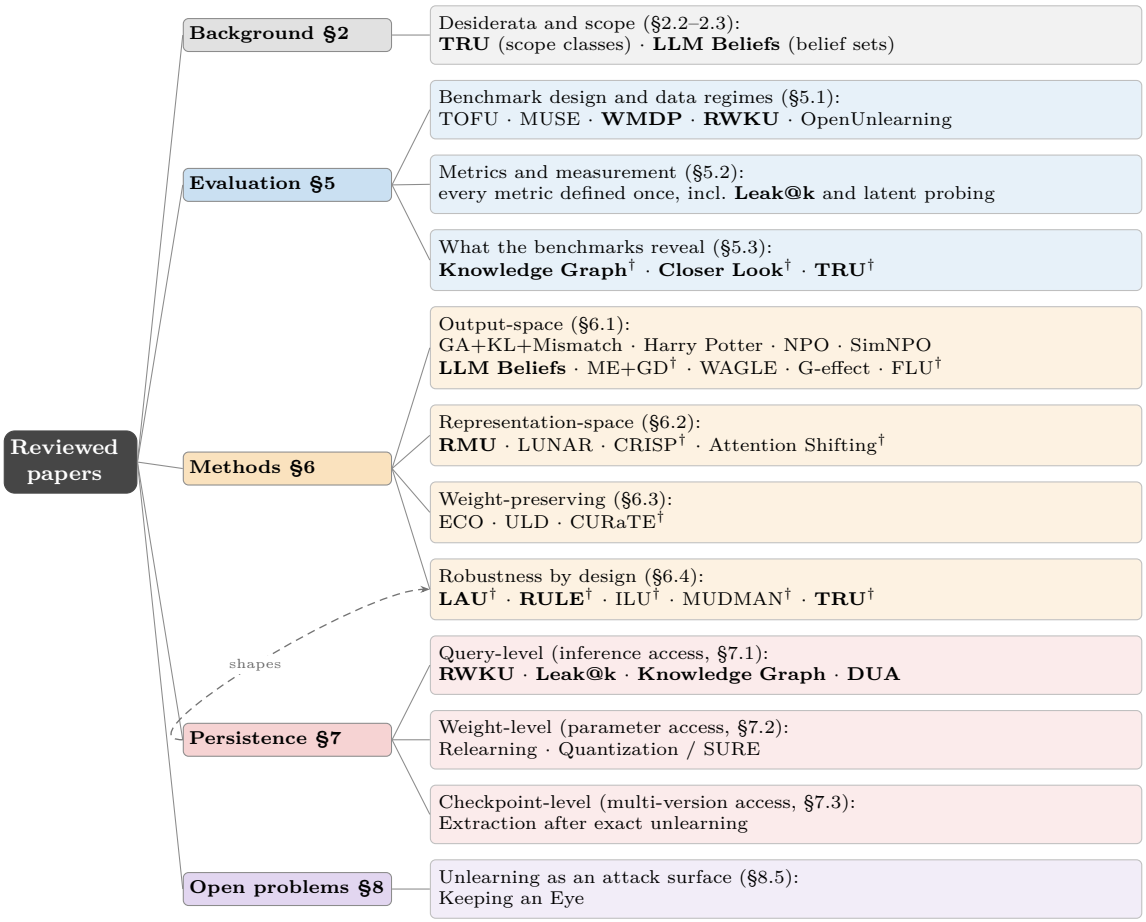

Figure 2: The reviewed papers, located by the section that develops them. **Bold** marks a paper with more than one home, so the same name appears in more than one list: WMDP supplies both a benchmark and the RMU method, RWKU both a benchmark and a persistence channel, and Leak@k a metric, a method and a channel. A dagger (†) marks a method paper that also contributes an evaluation metric. The dashed edge records the one non-hierarchical dependency: the recovery channels of Section 7 now shape the robustness-by-design methods of Section 6.4, so verification feeds execution as well as following it.

purpose: both inject the target, so neither can say what removal costs when it was absorbed at pre-training scale.

The other regime gives up that control to buy a knowledge origin nobody chose. WMDP and RWKU both target knowledge absorbed during pre-training (distributed across orders of magnitude more parameters and training steps than a fine-tuning injection, and not attributable to any one document), and they reach it by structurally different routes. WMDP goes through expertise. Li et al. (2024) commission four-choice questions in biology, cybersecurity and chemistry, each vetted by at least two experts from different organizations, targeting operational know-how rather than general awareness, with a security review to ensure the benchmark is not itself a hazard (counts in Appendix A.5). The authors test the benchmark's validity as a proxy against 122 biosecurity questions withheld during construction, and state separately that the questions stand in for hazardous knowledge in biosecurity and cybersecurity but *not* in chemistry. They are explicit about the limit: "models with a high score on WMDP are not necessarily unsafe." RWKU goes through breadth of probing instead. Jin et al. (2024) take 200 real public figures from Wikipedia, admitting a target only where models demonstrably memorize it, and withhold both the forget corpus and the retain corpus, so a method receives the target and the original model and nothing else. Around each target sit multi-level forget probes, nine adversarial probe types (Appendix A.1), and a neighbor set (Section 5.2); methods are

Table 1: Comparison of evaluation frameworks along key design dimensions. The last row is the one that matters for reading Sections 6 and 7: it records which instrument each later claim is actually reported against, so that an unlearning result and the conditions it was measured under can be held together. OpenUnlearning is a platform rather than a benchmark: it supplies no corpus of its own and re-runs prior methods under one standardized setup.

	TOFU	**MUSE**	**WMDP**	**RWKU**	**OpenUnlearning**
Domain	Fictitious authors	News; Harry Potter	Bio, Cyber, Chem	200 real people	Hosts existing benchmarks
Data type	Synthetic (GPT-4)	Real corpora	Expert-authored MCQ	Real (Wikipedia)	None of its own
Knowledge regime	Fine-tuning	Fine-tuning	Pre-training	Pre-training	Inherited from host
Metric philosophy	Distributional (KS)	Decomposed (C1–C6)	Operational (MCQ)	Multi-probe + adv.	Standardization; ranking sensitivity
Methods tested	GA, GD, KL, PO	GA, NPO, TV, etc.	GA, GradDiff, RMU	GA, NPO, DPO, etc.	Prior methods, one setup
Results reported against it	NPO, SimNPO, ECO, ULD, AS, LUNAR, RULE	RULE; GA and NPO variants	RMU, ECO, CRISP, TRU	Query-reformulation channel; partial-layer study	Ranking instability (§5.3)

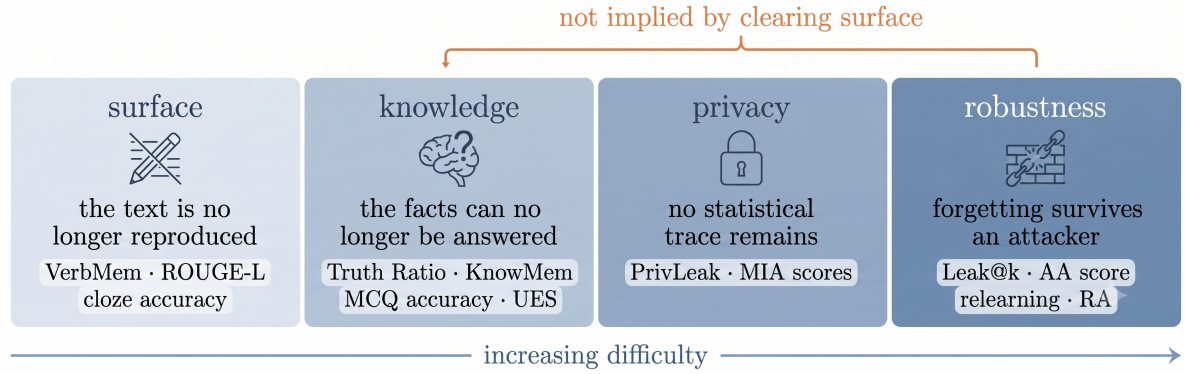

Figure 3: The forgetting hierarchy. Evaluation metrics are organized by the level of forgetting they assess, from surface-level verbatim reproduction through knowledge-level factual recall and privacy-level statistical indistinguishability to robustness-level resistance under adversarial or reformulated queries. Satisfying a weaker level does not guarantee a stronger one.

run under full fine-tuning, partial-layer fine-tuning and LoRA (Hu et al., 2022), and the partial-layer study reports that fine-tuning the early layers unlearns better without disturbing neighbor knowledge. Results are averaged over 100 single targets, with a separate batch-target experiment at sizes 10 through 50. Between them the pair covers the pre-training regime from two sides (Table 1). No benchmark holds the target itself fixed, so how much harder pre-training knowledge is to remove than fine-tuned knowledge remains unmeasured.

5.2 Metrics and Measurement

Whichever regime a benchmark works in, its verdict is delivered by a metric, and the metrics do not agree. We organize them by the level of forgetting they test, from surface reproduction to resistance under pressure, then by what they record of the damage, and finally by what falls outside it (Figure 3). Each is defined here once; later sections name them only.

5.2.1 Surface- and Knowledge-Level Metrics

VerbMem, ROUGE-L and cloze accuracy. Starting with the shallowest level: whether the string comes back. Three benchmarks ask it in three surface forms. Shi et al. (2024) define *verbatim memorization* (C1) through continuation: the model is prompted with the first l tokens of a forget-set sequence, and

the ROUGE-L overlap between its continuation and the true one is averaged over $\mathcal{D}_f$, lower being less reproduction. TOFU applies the same overlap in a QA format, between generated and ground-truth answers. RWKU narrows the surface to a single token: a *cloze* probe has the model complete a factual sentence about the unlearning target, scored by exact match against the ground-truth token. The three differ only in how much of the surface they hold fixed; aggregation formulas are in Appendix A.6.

Truth Ratio and Forget Quality. Matching the string is the weakest test, so the next family asks about the distribution instead. TOFU's *Truth Ratio* is the primitive: for each QA pair it compares the mean length-normalized probability of a set $\mathcal{A}_{\text{pert}}$ of factually incorrect perturbations $\hat{a}$ against that of a *paraphrase* $\tilde{a}$ of the correct answer,

$$R_{\text{truth}} = \frac{\frac{1}{|\mathcal{A}_{\text{pert}}|}\sum_{\hat{a}\in\mathcal{A}_{\text{pert}}} P_\theta(\hat{a} \mid q)^{1/|\hat{a}|}}{P_\theta(\tilde{a} \mid q)^{1/|\tilde{a}|}} \tag{4}$$

so a ratio above 1 marks a forgotten fact. Both departures matter: the fine-tuned phrasing carries inflated probability, so the denominator is a paraphrase, not the ground truth; length normalization keeps unequal answers comparable. *Forget Quality* aggregates it against the retrained oracle, as the p-value of a two-sample Kolmogorov–Smirnov test between the truth-ratio samples of θ_u and θ_r:

$$\text{Forget Quality} = p\text{-value of KS}\big(\{r_i^u\},\ \{r_i^r\}\big), \quad D_{n,m} = \sup_x \big|F_n(x) - F_m(x)\big| \tag{5}$$

A p-value near 1 indicates indistinguishability, making Forget Quality the most stringent criterion in this survey: a match in probability space, not on an aggregate. The *Inverted Truth Ratio* is an orientation convention on the same primitive (Li et al., 2026). All three read probabilities, none generated text: the probability–generation gap of Section 5.3.

KnowMem, MCQ accuracy and the Unlearning Effectiveness Score. Distributional tests still read one surface form, where the question is whether the knowledge is usable. Shi et al. (2024) define *knowledge memorization* (C2) as ROUGE overlap on QA pairs derived from the forget corpus, capturing answers knowable only from $\mathcal{D}_f$; it is stronger than VerbMem, since a model can paraphrase what it will not reproduce. WMDP scores four-choice accuracy, where 25% is random chance, testing operational capability rather than declarative recall. Both probe one hop. Wei et al. (2025)'s *Unlearning Effectiveness Score* closes that gap: an external judge rates how far a target triple remains inferable from the neighboring triples the model still asserts, normalized against the pre-unlearning model.

5.2.2 Privacy- and Robustness-Level Metrics

PrivLeak. All three ask what the model says; the next two ask what can be got out of it. Shi et al. (2024) define *privacy leakage* (C3) through a Min-k% Prob membership-inference attack (Shi et al., 2023) discriminating $\mathcal{D}_f$ from a holdout set, reporting the unlearned model's AUC as a normalized deviation from the retrained model's, $(\text{AUC}(f_u) - \text{AUC}(f_r))/\text{AUC}(f_r)$. Near zero is success. Their convention is that a large *positive* value marks *over*-unlearning and a large negative one *under*-unlearning. Both deviations are detectable: residual membership signal on one side, and on the other an atypically high loss on data a retrained model would never have seen. RWKU supplies four attacks of its own (LOSS, Zlib entropy and two Min-k% variants), and reports raw LOSS scores.

AA score and Leak@k. Membership leakage is measured with the weights fixed and the query fixed; relax the query and a second axis opens, ordered by how much effort the adversary is assumed to spend. RWKU's *Adversarial Attack* (AA) score assumes the least: nine hand-designed probe types reformulate each forget-set query (among them prefix injection, affirmative suffix appending, multiple-choice reformatting and reverse querying), and the score is the average accuracy across them, lower being more robust (Jin et al., 2024). Reisizadeh et al. (2026) assume an adversary who leaves the query alone and samples instead: given a query q and target answer a, *Leak@k* is the expected best of k independent generations,

$$\text{Leak@k}(q) = \mathbb{E}_{\mathbf{y}\sim P_\theta(\cdot|q)}\left[\max_{i\in\{1,\ldots,k\}} M(y_i, a)\right] \tag{6}$$

for an answer-matching metric M. Leak@k is not calibrated against zero. On TOFU the retrained oracle itself scores 0.622 at $k = 128$, so a Leak@k figure is a comparison against a reference that leaks, not an absolute quantity of leakage.

Relearning robustness and robust accuracy. And relax the weights instead, and a third opens: whether forgetting survives further training. Two protocols measure durability under relearning on forget-distribution data, and hold opposite things fixed. Sondej et al. (2025) fix the relearning stage (a fine-tuning budget held constant across the methods compared) and tune the *unlearning* hyperparameters by automated search against the post-relearning objective, so the reported forget-set loss is each method's best attainable irreversibility. Guo et al. (2024) fix the method and report a *relearning-recovery percentage*, replicating an earlier protocol: after retraining on half the forget set's labels, the fraction of forgotten accuracy recovered on the held-out half. Relearning is not the only way weights move; Wang et al. (2025)'s *robust accuracy* (RA) averages forgetting across a fine-tuning trajectory on an unrelated task.

5.2.3 Utility and Entanglement Metrics

Model Utility and neighbor preservation. Every one of these measures what was removed; none measures what was broken. Utility metrics run coarse to sharp. General-ability suites such as MMLU (Hendrycks et al., 2021) ask only whether the model still works (Li et al., 2024; Jin et al., 2024). TOFU's *Model Utility* is sharper: the harmonic mean of nine rescaled values, three per-sample metrics (ROUGE-L, answer probability, Truth Ratio) on each of three non-forget subsets, so that damage to one cannot average out. Sharpest is RWKU's *neighbor* set, which scores accuracy on probes about entities associated with the target and is the only measure here of entanglement rather than general damage.

5.2.4 Metrics Outside the Behavioral Hierarchy

Latent probes, judges and procedural metrics. Three metric families sit outside this behavioral hierarchy altogether. Guo et al. (2024) train a linear probe on residual-stream activations to predict the forgotten attribute; *latent-probe accuracy* is the rate at which it remains decodable after unlearning. That is the one instrument here that separates suppressed expression from altered representation. Ashuach et al. (2026) read it at the level of SAE features (Huben et al., 2024). A second judges generations rather than probabilities: Liao et al. (2026) score six judge dimensions split between unlearning quality and retention quality (Appendix A.3). Within unlearning quality, relevance measures how far a response *avoids* reproducing the target. Leng et al. (2025) add an Entailment Score, Tan et al. (2026) reproduction and hallucination rates. A third measures the procedure: MUSE's scalability as the forget set grows (C5) and sustainability across successive requests (C6), and Bae et al. (2026)'s unlearning time and inference overhead.

5.3 What the Benchmarks Reveal

Applied to the same methods, these instruments produce one result they all agree on and several they do not. The agreement is a negative: on none of them does any method satisfy all criteria at once. Maini et al. (2024) report that all four baselines lower Model Utility, and that methods move vertically in the Forget Quality–Model Utility plane without reaching the retrained oracle. Shi et al. (2024) find that most algorithms prevent verbatim and knowledge memorization to varying degrees, but that "only one algorithm does not lead to severe privacy leakage." Li et al. (2024) report that RMU drops WMDP-Bio and WMDP-Cyber to near random while holding MMLU, where the other baselines they test cannot drop those scores without crippling MMLU. Jin et al. (2024) find efficacy and locality impossible to balance at once: unlearning a target has side effects on neighboring knowledge and on truthfulness and fluency. The cost is not marginal: across MUSE's eight methods, *all* compromise utility by 24.2%–100%, three of them completely. What the four do not agree on is who fails least: re-running the same methods under one standardized setup, Kurmanji et al. (2025) find that the ordering among them shifts with the hyperparameters, the tuning checkpoint and the ranking scheme chosen, so reported numbers do not compare across papers (Appendix A.2). And every one of these results is a falsification. An instrument that reports a failure has told you something; no benchmark here establishes that a method which passes has removed anything.

The agreement is sharper than a shared failure: the difficulty is ordered. MUSE scores its criteria separately, and the scores fall in a consistent sequence. Most gradient-based methods drive VerbMem to near-retrain levels, so surface forgetting (C1) is cheap. Knowledge memorization (C2) resists longer, because a model that no longer reproduces a passage can still answer questions about it, and privacy leakage (C3) resists longest of the three. We read those three results as a difficulty ordering, and Sections 6 and 7 reason from it as one. MUSE states the criteria, not the ranking. At the top of the ordering forgetting and privacy come apart entirely. On NEWS, GA and NPO reach VerbMem 0.0 and KnowMem 0.0 (nothing verbatim, nothing recalled) and still score PrivLeak 5.2 and 24.4 (Shi et al., 2024). Under MUSE's convention those positive values mark *over*-unlearning, and that is the point worth holding: a model can be identified as unlearned by having tried too hard, because a retrained model would never assign atypically high loss to data it simply never saw.

That ordering is one the instruments can only report if they are measuring what they claim, and five results show they are not. Order them by how far past the instrument's own reading you have to go. Not at all, first: in an appendix Shi et al. (2024) give 95% bootstrap intervals for every C1, C2 and utility cell, and on NEWS `NPO_GDR`'s KnowMem interval is 54.6 [47.5, 61.5] against a retrained 33.1 [26.8, 39.5]. These are intervals the field quotes as point estimates. Permute the options next: Liao et al. (2026) report, also in an appendix, that reordering which choice is correct raises GradDiff's WMDP unlearning score **from 76.0 to 100**, because the unlearned model emits near-identical gibberish whose flat distribution favours option A. Read the generation rather than the probability: Leng et al. (2025) show that probability-based metrics report success on TOFU while the generated text still entails the original answer, and name the quantity that catches it, an Entailment Score. Reword the question: Jin et al. (2024) find that prefix injection, affirmative suffix, multiple-choice and reverse-query attacks effectively elicit knowledge the standard probes call gone. Never ask it at all: Wei et al. (2025) score a target against the supporting subgraph the model still asserts rather than as an isolated triple; in most settings the Unlearning Effectiveness Score falls by at least 20%, over half by more than 30%. Their second result closes the loop: where local utility is preserved the supporting structure survives and carries the inference, so the settings where that leakage disappears are the settings where the model has been broken. One result cuts the other way: Reisizadeh et al. (2026) find method rankings preserved as temperature, top-p and k vary. It sharpens the picture rather than softening it: rankings are stable under decoding and unstable under the choices above, so the instability is not noise but a property of what the field leaves unstandardized. No one of these five papers states the general claim, and two left theirs in an appendix. Read together they say it: a certificate of forgetting is not evidence that the forgetting happened.

Behind every one of these failures is the same mechanism. Removing one fact disturbs its neighbors: TOFU's fictional-author unlearning degrades performance on real authors and world facts (Maini et al., 2024), and RWKU records side effects on neighboring knowledge as a headline finding (Jin et al., 2024). Neither benchmark says why. The nearest mechanistic account is Guo et al. (2024)'s, which localizes factual recall to specific lookup components: a factual-recall editing result on sports facts and CounterFact, and one the benchmarks' own evidence does not reach. Entanglement scales with the request. On TOFU the failure at 1% is that no baseline crosses the 0.05 p-value threshold at all, and above 1% it becomes severe utility loss: two failures, not one widening gap. On RWKU, GA collapses at a batch-target size of 30, while refusal tuning stays stable. The methods of Section 6 were optimized against these instruments, which are now defined and known to overestimate what they certify.

6 Unlearning Methods

What separates one unlearning method from another falls on two axes: *where* the intervention lands, and *how* the optimization is stopped from destroying the model it edits. Nearly all of them modify weights with a small number of gradient steps on $\mathcal{D}_f$, and the first axis sorts those into three families. Output-space methods (Section 6.1) act on the distribution $P_\theta(y \mid x)$ itself, representation-space methods (Section 6.2) act on the hidden activations that produce it, and weight-preserving methods (Section 6.3) leave θ untouched and intervene at inference time instead. A fourth family (Section 6.4) inverts the question, asking not where to intervene but what the result has to survive, and builds resistance to a named recovery channel into the procedure itself. The second axis is the one this section follows from family to family, because the answer to

it turns out to be shared: every one of them arrives at a bounded optimization. Table 2 records both axes for every method named below, with the outcome each reports and the condition it was measured under.

6.1 Output-Space Methods: From Gradient Ascent to Bounded Objectives

Gradient ascent. Begin where the field began, with the objective everything else was built to repair. Write $\mathcal{L}(\mathcal{D};\theta)$ for the negative log-likelihood of $\mathcal{D}$ under θ. Gradient ascent (GA) reverses training by maximizing that prediction loss on the forget set, equivalently by descending on its negation:

$$\mathcal{L}_{\mathrm{GA}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\big[\log P_\theta(y \mid x)\big] = -\mathcal{L}(\mathcal{D}_f;\theta) \tag{7}$$

It is the cheapest thing that could work (one term, no retain data, no auxiliary model), and it is unbounded. The prediction loss has no upper limit, so nothing in the objective marks the point at which the forget set has been forgotten. Zhang et al. (2024) locate the consequence in the gradient itself: its scale does not shrink as $P_\theta(y \mid x)$ falls, so the parameters keep moving long after the forget-set probability is negligible and the model "diverges quickly from the initial model." What that costs on the benchmarks is in Section 5.3. The qualitative failure is separate from the numerical one, and stopping early does not repair it. GA is *untargeted* in the sense of Leng et al. (2025): it specifies no output for the forget set, only an output to move away from, and the displaced probability mass still has to land somewhere. Ji et al. (2024) print what that looks like when it lands badly: a GA-unlearned model answering "Sorry Christmas Christmas Christmas…" A repair therefore owes two things at once: a bound, so the optimization has a place to stop, and a target, so the mass has a place to go.

External regularization. The first repairs left the objective alone and supplied it with a target from outside. Eldan & Russinovich (2023) give the first effective approximate unlearning technique for a generative LLM: train a *reinforced* model further on the target corpus, compare its logits against the baseline model's to find the tokens that corpus is responsible for, replace the idiosyncratic expressions and entity names among them with generic counterparts taken from the model's own predictions, and fine-tune on the relabelled text as ordinary next-token prediction. No divergence penalty appears anywhere in the procedure; what is external is the reinforced model, which supplies the target the loss does not. The authors note that a corpus this dense in distinctive names "may have abetted" the strategy, and the relabelling itself runs on GPT-4 (OpenAI, 2023). Yao et al. (2024) build directly on that label translation and give it a general form, an update rule with three terms they name Unlearn Harm, Random Mismatch and Maintain Performance. Random Mismatch is the one that supplies the target: it keeps the forget prompt and attaches a response drawn at random from unrelated normal data, training the model *toward* that response rather than only away from the memorized one. Maintain Performance anchors predictions on a normal set through a KL penalty, and that set has to match the forget set's format (TruthfulQA for a question–answer forget set), which is a stronger requirement than any general corpus. From negative samples alone the method reports better alignment than RLHF at 2% of its computational time.

Bounding the objective from inside. Both repairs work, and both are superseded once the bound is moved inside the loss instead of being bolted to the outside of it. Zhang et al. (2024) take the DPO objective (Rafailov et al., 2023) and delete its positive term, since in unlearning there are no preferred responses, only responses to be suppressed. What remains is Negative Preference Optimization (NPO):

$$\mathcal{L}_{\mathrm{NPO}}(\theta) = -\frac{2}{\beta}\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\log\sigma\left(-\beta\log\frac{P_\theta(y \mid x)}{P_{\mathrm{ref}}(y \mid x)}\right)\right] \tag{8}$$

where P_{ref} is the frozen original model and $\beta > 0$ the inverse temperature. The forget-set term now sits inside a log-sigmoid, and the loss rewrites as $(2/\beta)\,\mathbb{E}[\log(1+(P_\theta/P_{\mathrm{ref}})^{\beta})]$, which is non-negative and approaches its floor as the forget-set probability goes to zero, so the objective supplies its own stopping point. In a standard logistic-regression setting (binary labels, a high-dimensional regime with $n_f \le d$), the paper's Theorem 2 bounds the weighted parameter-space distance $\|\theta^{(t)} - \theta_{\mathrm{init}}\|_{X^\top X}$ travelled after t steps, and the bound grows as $O(t)$ under GA against $O(\log t)$ under NPO; in the authors' own summary, NPO "diverges exponentially slower than GA in a simple setting." The quantity is a distance in parameters rather than between output

distributions, and the model is a logistic classifier rather than a language model. The mechanism sits in the gradient of the objective itself: NPO's is GA's scaled by an adaptive weight $W_\theta(x, y)$, and on an example the model has already unlearned, $P_\theta(y \mid x) \ll P_{\text{ref}}(y \mid x)$ drives $W_\theta \ll 1$, so the NPO gradient's norm drops far below GA's and NPO "could diverge much slower than GA." GA has no such factor and keeps pushing at full strength.

The bound arrives attached to a reference model, and two papers show it need not be. Zhao et al. (2025) name two defects in that attachment. First, the reference model allocates unlearning power unevenly: a strongly memorized example has high P_{ref} and so receives less first-order optimization power than its difficulty warrants. Second, because θ starts as a copy of the reference model, $W_\theta \approx 1$ for everything at initialization, so "NPO behaves similarly to GA in the early stage of unlearning", which is where the utility damage is done. SimNPO replaces the log-ratio with a length-normalized log-probability, adopting the reference-free SimPO as NPO adopted DPO:

$$\mathcal{L}_{\text{SimNPO}}(\theta) = -\frac{2}{\beta}\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\log\sigma\left(-\frac{\beta}{|y|}\,\log P_\theta(y \mid x)\right)\right] \tag{9}$$

The display omits the reward margin γ, set to zero to align with NPO and reported to accelerate the utility drop when raised. The $1/|y|$ carried into the gradient makes allocation deliberately length-*dependent*: less optimization power to longer responses. The bound survives without the reference model: the loss is bounded below by zero on its own, and the gain comes from smoothing the gradient weight. Leng et al. (2025) reach the same independence from the retain side, against a failure regularization does not prevent: the unlearned model becomes excessively ignorant, refusing most retain-set questions. Maximize entropy (ME) replaces the forget loss with the KL divergence from each next-token distribution to the uniform distribution over the vocabulary, supplying the target GA never had; answer preservation (AP) replaces the retain loss with a preference term that lowers the probability of the rejection template while maintaining that of the original answer, and needs no reference model either. Three unrelated routes (length-normalized sigmoid, fixed uniform target, reference-free preference term) reach one place: a bounded objective with no second model holding it there.

What the bounded objectives leave behind. Bounded or not, every one of these objectives moves probability mass rather than removing it. Li et al. (2026) name the consequence the *squeezing effect*: the conditional probabilities for a given input sum to one, so lowering the likelihood of the target response necessarily raises the likelihood of some alternative, and the mass lands in the high-likelihood region beside it, on semantically related rephrasings of the target. What the automated metrics read is the probability of the target string, which has fallen; what the model emits is the rephrasing, which they never score. The paper calls the result *spurious unlearning* and demonstrates it on GA and NPO under TOFU's 10% forget set with Llama 3.2 1B and greedy decoding. Its answer folds the model's own beliefs (the top-k next-token set locally, high-confidence sampled generations globally) back into the objective: BS-T replaces the one-hot target with a soft target mixing in that top-k distribution, and BS-S samples completions from the model being unlearned and minimizes a convex combination of the losses on the original and the augmented forget sets (Appendix A.15).

A gradient-level diagnostic says why, and separates the objectives that damage the model from those that do not. Liu et al. (2025a) define the gradient effect (G-effect) of an unlearning objective as a scalar, the inner product between the gradient of a risk metric and the gradient of the unlearning loss at the current parameters: positive means the update improves that risk metric, negative means it harms it. Read on the forget set the same quantity gives the *unlearning* G-effect, read on its complement the *retaining* G-effect, and the two are wanted in opposite directions: notably negative for the first, since opposing the forget-set risk is what removal requires, and non-negative for the second. So GA's characteristic failure is not insufficient unlearning but excessive unlearning, extreme negative values of the unlearning G-effect, while NPO's appears as a negative retaining G-effect. The paper builds weighted GA (WGA) and token-wise NPO (TNPO) on the diagnosis.

Fact-lookup localization (FLU) and weight attribution (WAGLE). If the damage is a matter of which parameters move, then restricting which parameters may move should be the remedy. Two methods

do it by criteria that range over different objects. WAGLE (Jia et al., 2024) ranks individual weights by an attribution score obtained in closed form from a bi-level formulation with unlearning above and retention below: $S_i \propto [\theta_o]_i[\nabla\mathcal{L}_f]_i - \gamma^{-1}[\nabla\mathcal{L}_r]_i[\nabla\mathcal{L}_f]_i$, whose second term (the interaction of retain and forget gradients) is what makes the criterion account for retention at all. Any objective is then run on the selected weights and improves its forget–retain trade-off on TOFU and WMDP, at a ratio greedy-searched per task; the worked example keeps 80%, so this is weight attribution rather than sparse surgery. Guo et al. (2024) rank components instead, by computational role rather than by any unlearning objective: fact-lookup unlearning (FLU) targets the intermediate feed-forward components that enrich a subject's representation with its attributes, against output-tracing (OT), which takes whatever most affects the output (Appendix A.13). Updating only the FLU components is more robust across input and output formats and "resists attempts to relearn" under an inherited protocol, a property no result has yet shown the weight-ranking criterion to buy. Probing shows the knowledge altered rather than rerouted around a fixed output path. That finding is the sharpest challenge in this corpus to the reading this survey advances, and it is a claim about where the change lands, not about what can still be got out: the paper runs no extraction attack, no relearning from public data, no comparison across checkpoints, and its scope is factual-recall editing on sports facts and CounterFact at Gemma and Llama scale. It shows the localization criterion to be a lever on robustness, which nothing in the corpus opposes, and leaves removal untested rather than answered.

6.2 Representation-Space Methods: Intervening at the Activation Level

Localization still edits weights against a loss defined on the output; the next family changes the target to the representation itself. Representation Misdirection for Unlearning (RMU), proposed alongside WMDP by Li et al. (2024), regresses the forget-set activations at an intermediate layer l, token by token, onto a fixed vector $c\,\mathbf{u}$:

$$\mathcal{L}_{\text{forget}} = \mathbb{E}_{x_f \sim \mathcal{D}_f}\left[\frac{1}{L_f}\sum_{t \in x_f}\left\|\mathbf{h}_l(t) - c\,\mathbf{u}\right\|_2^2\right] \tag{10}$$

with $\mathbf{u}$ a random unit vector, entries uniform on $[0, 1)$ and fixed throughout training, c a deterministic scaling hyperparameter, and L_f the token count of x_f. A parallel term pins the retain-set activations to the frozen original model's,

$$\mathcal{L}_{\text{retain}} = \mathbb{E}_{x_r \sim \mathcal{D}_r}\left[\frac{1}{L_r}\sum_{t \in x_r}\left\|\mathbf{h}_l(t) - \mathbf{h}_l^{\text{orig}}(t)\right\|_2^2\right] \tag{11}$$

and the two combine as $\mathcal{L}_{\text{RMU}} = \mathcal{L}_{\text{forget}} + \alpha\,\mathcal{L}_{\text{retain}}$. Both halves of the target do work, and the paper leads with the one easiest to overlook: the update "changes direction and scales up the norm" of activations on hazardous data, because "increasing the norm of the model's activations on hazardous data in earlier layers makes it difficult for later layers to process the activations." That is a claim about disrupting downstream processing, not about removing what the earlier layers still encode. Because the regression target does not move, the optimization has nowhere to run. On WMDP, RMU drives hazardous-domain accuracy to near the 25% random baseline while preserving nearly all MMLU, and its robustness claim covers linear probes and adversarial queries. It does not cover fine-tuning. The paper's own Appendix B.6 reports that "RMU is unable to prevent finetuning from recovering performance," and the authors conclude in their own voice that "RMU perhaps obfuscates knowledge more than unlearns it." Layer ranges, coreset and compute are in Appendix A.14.

A second design reaches the same redirection without RMU's multi-term objective at all. Mao et al. (2025) redirect the representations of unlearned data into activation regions that "express its inability to answer," using a *single-term* loss ("in contrast to many prior approaches that rely on multi-term objectives") that reduces the intervention to one down-projection matrix for which a convergent closed-form solution exists. RMU is a baseline LUNAR beats rather than a design it modifies, and the gains are on both axes at once: 2.9×–11.7× on the combined Deviation Score, and roughly 20× in efficiency. Concentrating the update in one matrix also carries the corpus's only claimed defence against a persistence channel, robustness to quantization attacks (Section 7.2). That claim rests on this one design feature and no one outside the paper has tested it.

Structures inside the layer. Both act on whole layers, where the knowledge is carried by structures inside them, and two methods name different structures. Ashuach et al. (2026) name feature directions. CRISP selects the sparse-autoencoder features strongly and frequently activated by the target corpus relative to a benign retain set, then fine-tunes through LoRA to suppress them on forget inputs while holding them steady on benign ones (Appendix A.16). It needs a pretrained SAE for the target model, and it targets the selectivity RMU and LUNAR lack: prior methods "impair performance on related but benign knowledge" and redirect the conversation "even in response to harmless questions." Its framing is mixed: the abstract says it removes knowledge, Figure 1 says it suppresses features. Its Limitations settle which: "residual information may persist in distributed representations, and robustness against adversarial extraction remains an open direction." Tan et al. (2026) name attention allocation instead. Attention Shifting scores token importance by the entropy change when a token is masked, then minimizes a convex combination of two KL divergences between attention distributions running in opposite directions: model to a reweighted target on the forget set, target back to model on the retain set (Appendix A.17). It places itself on the aggressive horn of its own dilemma, against conservative methods that preserve utility but risk hallucinating, ULD among them.

Taken together the two groups answer the same question by opposite routes and arrive at the same place: a bounded optimization, and an intervention that redirects expression rather than deleting it. What no one has measured is whether redirection leaves the statistical traces a membership-inference attack detects. RMU, LUNAR, CRISP and AS are scored on capability, overlap and utility; none has been run against PrivLeak or any other membership-inference measure, and the one paper in the family that raises the risk at all, Mao et al. (2025), argues that prescribing a fixed refusal makes membership easier to infer, and measures nothing. Where the coupling has been measured, Wei et al. (2025) find inferential leakage suppressed only in the settings where the model has been damaged (Section 5.3). On this evidence the untested axis is not a gap in coverage but a gap of exactly the kind the preceding results predict, and Section 7 takes up what the evaluations assumed.

6.3 Weight-Preserving Methods: Parametric Versus Behavioral Unlearning

Both routes still change the weights, and a third family declines to. Bae et al. (2026)'s CURaTE and Liu et al. (2024)'s ECO both leave θ frozen and gate the query, and disagree about what the gate should then do. CURaTE embeds each deletion request into a vector store and routes incoming prompts by cosine similarity against a threshold δ: a match receives a coherent refusal, everything else reaches the unmodified model. Nothing is retrained, so routing costs sub-millisecond latency and no retained capability, however many requests arrive (Appendix A.18). ECO gates the same way and rejects the refusal. Its classifier triggers a corruption of the input rather than a replacement of the output, and it is narrow: only "the first dimension of each token's embedding," at a strength σ tuned by a zeroth-order search not to maximize forget-set loss but to match a surrogate of what a retained model would score. The objective carries no retain-set term, and retention is preserved by the gate instead (Appendix A.7). The reason is ECO's own: a template response "violates the weak unlearning objective... because a retained model is highly unlikely to give template responses." Its zero-out variant, ECO-ZO, reaches Forget Quality comparable to the retrained oracle in all six TOFU settings, at "zero sacrifice in model utility." No gradient passes through the model, so ECO scales without per-model retraining, though it needs embedding-layer access that a token-only API does not give. Both then inherit their gate's recall: ECO misses 7.52% of forget queries under jailbreak-like prefixes, 1.75% after augmentation; CURaTE's own appendix reports 41% recall under payload splitting.

Gating the output instead (ULD). Gating the input requires knowing which inputs to gate; gating the output does not. Ji et al. (2024) train a small assistant with the forget and retain objectives reversed, $\min_\phi \mathcal{L}_f(\phi) - \beta\,\mathcal{L}_r(\phi)$, so that it memorizes the forget set and is driven toward a uniform distribution on the retain set. They then subtract its logits from the frozen target's at decode time, $\mathbf{l} - \alpha\,\mathbf{l}_a$. This is the forget–retain asymmetry of Section 2 used as a construction: the tractable objective, memorizing a bounded forget set, goes to the assistant, and forgetting is obtained by subtraction. What bounds the procedure is that flip in direction, not the frozen target: the assistant is itself trained by gradient descent, and boundedness comes from minimizing a cross-entropy bounded below, toward a uniform target (Appendix A.9). The costs are specific. The assistant runs alongside the target at inference and trains for ten epochs on TOFU before deployment, so the family's zero-latency property is CURaTE's, not ULD's. Retain ROUGE-L falls to

98.3, 93.4 and 85.9 at forget-set sizes of 1%, 5% and 10%, and Forget Quality to 0.99, 0.73 and 0.48. The near-perfect figure is the 1% setting alone.

What all three buy is the same, and it is less than it appears. Expression is suppressed while the weights are untouched, so the knowledge stays fully encoded and an adversary who can read the weights bypasses the intervention entirely. Liu et al. (2024) say so themselves, naming their guarantee *weak unlearning* under a declared gray-box threat model: users see completions and logits, never weights. That is not a weakness of the family: it is the only place in this corpus where a guarantee is stated relative to a declared adversary, which is what Section 8 argues every method should do.

6.4 Designing for Robustness: Unlearning That Resists Recovery

Each family so far is defined by where it intervenes; the last is defined by what it expects to be attacked with. These methods make robustness an objective of the procedure rather than a property measured once unlearning is finished, so the design question moves from how far to push the forget-set probability down to how to make the reduction survive. Four pressures have been named (an adversarial query, a stochastic decoding budget, a downstream fine-tune on unrelated data, a deliberate relearning attempt), and each has a method built against it. None replaces the three families above: they wrap output-space objectives, compose with representation-space ones, or restrict which weights may move. What individuates them is the failure mode they anticipate rather than the site at which they act, and those failure modes are the subject of Section 7.

Adversaries that arrive at inference. The first two anticipate an adversary who arrives at inference, and both answer by simulating that adversary inside the training loop. Yuan et al. (2024) show that an optimized adversarial suffix resurfaces forgotten knowledge, and reply with Latent Adversarial Unlearning (LAU), a saddle-point formulation in the paper's own words:

$$\max_{\theta} \; \mathbb{E}_{(x_f, y_f) \sim \mathcal{D}_f} \min_{\|\delta\| \le \kappa} \; \mathcal{L}\big(\theta, \delta;\, x_f, y_f\big) \tag{12}$$

The inner minimization perturbs the residual stream at a chosen layer within a norm bound κ to reactivate the knowledge, standing in for what an input attack would have done. The outer step hardens against it. Perturbing latents rather than discrete tokens keeps the inner attack tractable (Appendix A.10). The variants are AdvGA and AdvNPO, and the headline (effectiveness improved by over 53.5% at under 11.6% of neighbouring knowledge) is AdvNPO's alone; AdvGA on LLaMA-3.1-8B costs 41.3%. Reisizadeh et al. (2026) simulate a different adversary, the decoder. Their RULE leaves the base objective untouched and augments the data it runs on: sample the current model k times per forget query under probabilistic decoding, admit every generation whose score clears a threshold into an augmented forget set, and re-run, three times over. The sampling budget enters through the data rather than through the loss (Appendix A.19). It suppresses leakage across $k \geq 64$ on TOFU and not on MUSE, which is the shape of the family: each hardens against the threat it could simulate.

The same wrapping strategy transfers to a threat that arrives *after* deployment rather than during inference, and transfers badly, because an inference-time adversary can be simulated inside the training loop while a future fine-tune cannot. Wang et al. (2025) treat each downstream task as an environment the unlearned solution should be invariant to, and borrow invariant risk minimization's practical penalty: the squared gradient of a loss on an unlearning-irrelevant fine-tuning set with respect to a scalar predictor held at unity, added to whichever base objective is in use (Appendix A.12). One unrelated fine-tuning set confers invariance that generalizes to unseen tasks, and ILU matches a meta-learning tamper-resistance baseline at a 63× efficiency advantage, lifting average robust accuracy from 0.42 to 0.65 on RMU and from 0.47 to 0.56 on NPO. What it does not buy is resistance to relearning: it loses there to the sharpness-aware baselines, its own conclusion lists relearning as future work, and its budget is 60 samples for one epoch. Sondej et al. (2025) take it from the other end, shaping the updates rather than the landscape. MUDMAN periodically trains an adversary on the forget set so dormant circuits reactivate and the unlearning gradient can reach them, normalizes that gradient as it decays, and applies an unlearning update to a weight only when its sign

agrees with the accumulated retain gradient's:

$$g_{\text{final}} = \begin{cases} g_u & \text{if } \operatorname{sign}(g_u) = \operatorname{sign}(g_r), \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Read against WAGLE's attribution mask (Section 6.1), the two restrict weights on opposite principles: WAGLE computes one mask before unlearning from a score combining both losses and keeps most of them, while MUDMAN recomputes its mask every step from sign agreement alone, and rejected the alternative on evidence, a tuned magnitude-percentile mask of WAGLE's kind having given no advantage at more tuning cost, and its own ablation putting most of the improvement over TAR on the mask rather than on the other two components (Appendix A.11). Scope is the caveat. MUDMAN's largest model is 1B parameters, an order of magnitude below every other method here, and its gains are measured against a stripped TAR baseline; its own appendix calls the second application a partial success where meta-learning does not help at all.

Specifying the behaviour instead. Those four train against a threat; the last specifies what the model should do instead. Liao et al. (2026) argue that gradient ascent is untargeted twice over: it fixes neither the *scope* of what is forgotten, so a paraphrase retrieves it, nor the *response*, so the model degenerates. Targeted Reasoning Unlearning uses a reasoning model to build an explicit target for each forget item (a trace that exposes the underlying knowledge, paired with a coherent refusal), and trains on those targets with a supervised loss beside a gradient-ascent term. Learning the trace is what buys the robustness: the model learns to recognize when a query falls inside the intended scope, so forgetting attaches to the equivalence class of Section 2 rather than a surface form, and survives cross-lingual queries, jailbreak prompts and relearning. That robustness is not inherited from the model that generated the targets, and the paper rules this out, varying three generators and two judges. The costs are a reasoning trace per forget item, a scope broad enough to start refusing benign queries, and an evidence base of WMDP-Bio alone: no baseline in the attack figure, relearning budgets of 15 and 5 samples, and retention specificity and logic below the un-unlearned model's.

Each of these five is defined relative to a threat it chose, and the choosing is the limit. LAU hardens against a latent adversary, RULE against a sampling budget, ILU against a fine-tune, MUDMAN against relearning, TRU against reformulation. None reports robustness across the set, and none addresses the channels that need no adversary at all, quantization or a second released checkpoint. Every family in this section reaches a bounded optimization; none has been shown to remove the knowledge from the weights. Section 7 drops the access assumptions the evaluations made.

7 The Persistence of Forgotten Knowledge

Each family of Section 6 bounds an objective and then reports what the evaluation can see. What the evaluation can see is narrow. Every framework in Section 5 grants the adversary exactly what the evaluator had (one model, one phrasing of the query, one decoding pass), and each result below relaxes one of those assumptions. They sort by how much access recovery requires: inference alone (Section 7.1), the weights (Section 7.2), or more than one released version of the model (Section 7.3). Figure 4 sets out the three levels.

7.1 Query-Level Recovery: Elicitation Under Fixed Weights

The weakest assumption to relax is the one every benchmark makes without noticing. Unlearning evaluations almost universally decode greedily and ask whether the top-1 continuation carries the forgotten content; deployment does not. Reisizadeh et al. (2026) sweep temperature and top-p jointly, with top-p the primary driver, and draw k independent generations per query. On TOFU and MUSE, and for RMU on WMDP, the probability of drawing at least one leaking generation rises sharply with k, the effect their Leak@k metric quantifies, while NPO on WMDP stays flat at every k, having over-forgotten. The scope of that word *reliably* is those three settings, not the method class. Nothing here is adversarial: the prompt is unchanged, the weights are unchanged, and the only thing that varies is how the model is sampled. The bound on what

Table 2: Comparison of unlearning methods along key dimensions. WAGLE and FLU are localization schemes, and LAU, ILU, MUDMAN, and RULE are training wrappers, each applied on top of a base objective rather than defining a standalone loss; their stability and retain-set requirements are largely inherited from that base. The *Robust to* column records the recovery condition against which a method reports robustness, drawn from adversarial querying, probabilistic decoding sampling, downstream fine-tuning, relearning, quantization, and rephrasings. A blank entry indicates no such claim.

Method	Space	Core mechanism	Stability	$\mathcal{D}_r$?	Robust to	Reported outcome
GA	Output	Ascend prediction loss on $\mathcal{D}_f$	Unbounded	No		Drifts from the initial model without stopping; collapses at RWKU batch-target size 30
GA+KL+ Mismatch	Output	GA + random mismatch + KL anchor	External reg.	Yes		Better alignment than RLHF at 2% of its compute, from negative samples alone
HP (generic-label translation)	Output	Relabel target tokens, then finetune	Base loss	No		Generic completions on target prompts; authors report residual wiki-level knowledge
NPO	Output	DPO-derived bounded loss	Adaptive weight	Opt.		$\|\theta^{(t)} - \theta_{\text{init}}\|$ grows $O(\log t)$ against GA's $O(t)$, in a logistic setting
SimNPO	Output	Ref-free bounded loss	Bounded below	No	Relearning	Length-*dependent* allocation via an explicit $1/\|y\|$; no reference model at all
BS-T / BS-S	Output	Soft token/seq. belief suppression	Bounded	Opt.	Spurious rephrasing	Counters squeezing in settings where target-only suppression scores as success
ME+GD	Output	KL to uniform in place of GA	Fixed target	Yes		Supplies the forget-set target GA lacks; paired with AP against over-refusal
WAGLE	Output	Bi-level weight attribution	Base loss	Base		Better forget–retain trade-off (TOFU, WMDP), at an 80% selection ratio
FLU	Output	Fact-lookup localization	Base loss	Yes	Adv. query, relearning	More robust across input/output formats; resists relearning
RMU	Repr.	Misdirect + scale up activation	MSE target	Yes	Adv. query	WMDP near 25% chance, MMLU ≈ preserved; finetuning recovers performance (App. B.6)
LUNAR	Repr.	Single down-projection redirection	Closed form	Yes	Quantization	2.9×–11.7× Deviation Score over baselines; ≈20× efficiency
CRISP	Repr.	SAE feature fine-tuning *(needs a pretrained SAE)*	MSE target	Yes	Concept recovery	Suppresses target features on WMDP-Bio/Cyber; paper concedes residual distributed information
AS	Repr.	Attention-KL shifting	Adapter reg.	Yes	Hallucination	Competitive ROUGE-L on TOFU; FQ below IHL and ULD, near GA
CURaTE	Behavioral	Real-time embedding matching	N/A (frozen)	No	Sequential updates	Sub-millisecond refusal routing; jailbreak results weaker than the framing suggests
ECO	Behavioral	Corrupt one embedding coordinate	N/A (frozen)	No		ECO-ZO: FQ ≈ oracle in all six TOFU settings; WMDP near-random at full MMLU
ULD	Behavioral	Assistant logit difference	Bounded	Yes		FQ 0.99 / 0.73 / 0.48 at 1 / 5 / 10%; retain ROUGE-L falls to 85.9
LAU (AdvGA/NPO)	Output	Latent adversarial training	Base loss	Base	Adv. query	AdvNPO headline 53.5% / 11.6%; AdvGA costs 41.3% of neighbor knowledge
RULE	Output	Iterative self-generation + forget-set augmentation	Base loss	Yes	Probabilistic sampling	Suppresses across $k \geq 64$ on TOFU; not on MUSE
ILU	Out./Repr.	Invariance (IRMv1) reg.	Base loss	Yes	Fine-tuning	63× efficiency over TAR; loses to SAM baselines on relearning
MUDMAN	Output	Meta-learn + disruption mask	Non-disruptive	Yes	Relearning	Beats a stripped TAR at ≤1B parameters; BeaverTails a "partial success"
TRU	Output	Reasoning target + GA	Coherent refusal	Yes	Adv. query, relearning	Relevance 0.04 (base) → 0.96 (GradAscent); WMDP-Bio only; retention below the un-unlearned model

this shows sits in the same table. The retrained oracle, the model the rest of this survey measures against, itself scores 0.622 at $k = 128$ on TOFU. A nonzero Leak@k is therefore not on its own evidence of failed unlearning, because the comparison is against a reference that leaks too.

Grant an adversary nothing but a rephrasing, and a second channel opens. Jin et al. (2024) run nine probe types against 200 unlearning targets on two models and report that four of them (prefix injection,

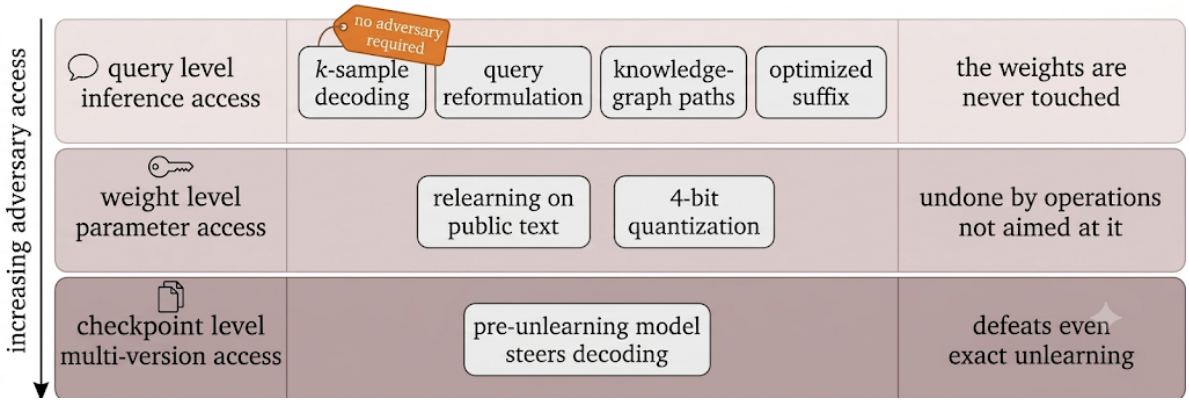

Figure 4: The three-tier persistence hierarchy. Knowledge that appears forgotten under standard evaluation can be resurfaced at higher tiers of adversarial access. Lower tiers require minimal access; higher tiers reveal deeper residual traces.

affirmative suffix, multiple choice and reverse query) effectively elicit unlearned knowledge from models that pass the standard cloze and question-answer probes. The direction of that finding matters: these are attacks that work, not knowledge that resists. Exposure is not uniform across methods. Because refusal tuning is fine-tuned on refusal data it achieves the best unlearning efficiency under adversarial attack in their evaluation, and NPO also shows some capacity to resist. The channel is also the oldest in the corpus. Eldan & Russinovich (2023) traced a small number of leaks in 2023: prompted for a list of fictional schools, their unlearned model still answered Hogwarts. They are explicit, though, that "none of these leaks reveals information that would necessitate reading the books," so what surfaced was the entity and not the content. Their next observation is worth more to this survey than the leak: their own evaluation, they write, "could potentially be blind to more adversarial means of extracting information." That is the shape of the channel. A rephrasing that recovers the answer falsifies a removal claim outright, while surviving nine probe types establishes nothing, because the probes are a finite hand-built sample of an open set. The authors who built the field's first such evaluation said so first.

Third, an adversary who brings outside structure. Wei et al. (2025) score a target triple against the supporting subgraph the model still asserts rather than in isolation; Section 5.3 gives the instrument and its numbers. It measures rather than attacks: an external judge rates inferability against the pre-unlearning model and a reference knowledge graph. But no single-hop probe reaches the residue it finds.

Last, an adversary who searches for the query instead of guessing it. Yuan et al. (2024) optimize a universal adversarial suffix that, appended to a probe, drives an unlearned model to reproduce forgotten content: an escalation from RWKU's hand-built probes to automated search, and one that scales across targets without manual effort. The consequential part is where the optimization runs. A suffix trained against the *original* model transfers, so the attacker recovers content "even without the direct access to the unlearned model itself"; what the attacker does need is white-box gradients on that original, an assumption open-weight releases satisfy. Recovery is strongest when the suffix is reused on the questions it was optimized against, weakens on new questions about the same target, and weakens further across targets, so the threat is sharpest where the adversary already knows what to ask. Scope is ten unlearned models on one benchmark. What the paper reports is a mean ROUGE-L recall between the elicited text and the forgotten answer, which the authors gloss as knowledge recovered in 55.2% of questions and render as 54% elsewhere in the same paper. It is not an attack-success rate, and no reading should treat it as a count of questions fully answered. The construction and generalization protocol are in Appendix A.22, and the defense it provoked, LAU, is in Section 6.4.

7.2 Weight-Level Recovery: Reversal Under Parameter Updates

All four of those channels leave the weights alone; grant the adversary the weights and two more open. Hu et al. (2025) fine-tune an unlearned model on a small set of public information about the forgotten topic (for a model unlearned on the Harry Potter novels, text covering the main characters, "which does not contain direct text from the novels"), and the forgotten facts return. The construction is the point. The obvious objection to any relearning result is that the relearn set simply re-teaches the answer, and the paper is built to rule that out: it criticizes a prior study whose relearn set could directly answer the evaluation queries, and keeps its own free of the source text (Appendix A.20). Its second adversary setting, in which the attacker holds a limited subset of the unlearning data, is not a clean recovery result and is not what is claimed here. What the clean setting shows is that unlearning left something to reactivate: the suppression is undone by data that never contained the suppressed content, which is hard to reconcile with the knowledge having been removed. The ILU framework of Section 6.4 responds to this channel directly, by enforcing invariance under fine-tuning updates.

Relearning at least intends to change the model; the next channel does not. Zhang et al. (2025) show that routine post-training quantization reverses unlearning: for methods that operate under a utility constraint, the unlearned model retains an average of 21% of the intended forgotten knowledge at full precision, and 83% of it after 4-bit quantization with standard techniques such as AWQ or GPTQ. The mechanism is arithmetic rather than adversarial. Unlearning under a utility constraint moves weights only slightly, by design, and 4-bit quantization bins are coarse enough that the move falls back inside the original bin, so rounding restores the pre-unlearning value. At 8 bits the bins are fine enough that it does not, and quantized models behave like full-precision ones: the failure belongs to aggressive quantization, not to quantization. The authors propose SURE, which folds rounding noise into the unlearning objective (Appendix A.8). One method in this survey claims a defense: Mao et al. (2025) argue that concentrating LUNAR's update in a single down-projection matrix makes it robust to precisely this attack (Section 6.2). That claim is the paper's own, untested by anyone else, and it covers one channel.

7.3 Checkpoint-Level Recovery: Extraction Across Model Versions

Both of those channels assume one set of weights, and deployment rarely produces only one. Wu et al. (2025) take the case a deletion request actually creates: a provider releases a model, receives the request, and releases the post-unlearning version, leaving an adversary holding both. Their attack generates from the post-unlearning model while using the pre-unlearning model as a reference to steer decoding, then restricts candidate tokens to those the pre-unlearning model rates highly (Appendix A.21). What is compared across the two versions is behavior rather than parameters, so checkpoint access can mean two logit APIs and not two weight files, and extraction success roughly doubles over prior baselines on MUSE, TOFU and WMDP. The setting is what makes this the most consequential result in the section. The paper's target is *exact* unlearning (retraining from scratch without the forget data). That model is not one more method: it is the retrained oracle, the reference every instrument in Section 5 measures against, from TOFU's Forget Quality to MUSE's PrivLeak. Under checkpoint access the gold standard itself leaks, and it leaks because of the unlearning, since a single released model gives the adversary nothing to compare against. A guarantee stated as indistinguishability from a retrained model inherits whatever that retrained model discloses, which is why the authors close on a threat model rather than a method: evaluation should account for adversarial access to prior checkpoints.

7.4 Synthesis: What Seven Channels License

Seven channels, and the question is what their existence licenses anyone to conclude. They are stochastic decoding (Reisizadeh et al., 2026), query reformulation (Jin et al., 2024), inferential paths through a knowledge graph (Wei et al., 2025), an optimized adversarial suffix (Yuan et al., 2024), benign relearning (Hu et al., 2025), post-training quantization (Zhang et al., 2025), and cross-checkpoint extraction (Wu et al., 2025). Those are seven papers, no two of which share a threat model, none of which treats the others as instances of the same thing, and no two of which were run on a common experimental setup. Taken together they carry an asymmetry that we read as the central fact about verification in this field. Each channel can

falsify a removal claim: one recovery under one channel settles the question, and settles it negatively. None of them can verify one. Surviving all seven says only that seven particular procedures failed to recover the content, and on this evidence the seven are not the whole set. Wu et al. (2025)'s channel operates outside the evaluation paradigm the other six inhabit, and Wei et al. (2025) state in their own Limitations that their measurement is a lower bound, since "real-world adversaries could potentially surpass these evaluators in inference power." Two cited papers, on independent grounds, say the enumeration is open. What follows is a restriction on what a removal claim can mean at all: a method that passes every published test has earned a negative result about the tests, not a positive one about the model.

If none of them can verify, certification has to become something other than a pass. What a benchmark issues today is a score under a single access model (inference only, one phrasing, greedy decoding, one released version), and none of the four frameworks of Section 5 tests any of the three levels above it. The gap is not only a research gap, because the channels track a deployment pipeline rather than an attack surface: models are quantized before serving, adapted to downstream tasks, released in successive versions, and sampled rather than decoded greedily. The conditions under which forgetting fails are routine operations, so a method certified under evaluation conditions and deployed under these has not been tested where it will run. Nor is the evidence assembled anywhere a practitioner could consult it: the persistence results are scattered across seven papers with seven experimental setups, and no standardized suite covers the pipeline end to end. That is the constraint Section 8 inherits, and it falls on certification rather than on optimization. The methods of Section 6 work as optimizations: they bound a loss, they trade forgetting against utility, and they improve on one another. It is the verification step that has no instrument, which is why the open problems that follow are dominated by measurement rather than by method design.

8 Open Problems and Future Directions

Sections 5–7 leave seven open problems and a proposal for what verification should become. Each is stated from evidence already reviewed rather than from what the field says it needs next.

8.1 Unified and Reliable Evaluation

The first problem is the one every other is measured through. No paper in this corpus reports a method on all four benchmarks of Section 5.1, and the platform built to end that fragmentation covers three of them (TOFU, MUSE and WMDP), leaving RWKU outside. Kurmanji et al. (2025) give the cause in their own terms: methods are not implemented across benchmarks, negative preference optimization is formulated differently on TOFU than on MUSE, and RMU was introduced for WMDP alone. Within any one benchmark the same fragmentation reorders who wins (Section 5.3). One axis is exempt: rankings hold as temperature, top-p and k vary. That exemption is what makes the instability specific rather than generic, since it belongs to the choices the field has not standardized.

Closing it needs expansion along two axes, and one coverage gap sits in neither. The axes are standardized setups that fix the experimental choices, and stress tests that ask whether forgetting survives sampling, inferential structure and the post-deployment procedures of Section 7. The gap is knowledge origin: TOFU and MUSE inject the forget corpus by fine-tuning while WMDP and RWKU operate on off-the-shelf models, and no benchmark measures the difference between removing what was fine-tuned in and what was pre-trained.

8.2 Genuine Forgetting Versus Obfuscation

And even a standardized instrument would not distinguish suppression from removal. Output-space objectives optimize the probability of the target string, a proxy for the quantity of interest, and the two come apart in both directions reviewed above: gradient ascent and negative preference optimization displace mass onto rephrasings the metrics never score (Section 6.1), while Ashuach et al. (2026) concede that weight editing toward concept erasure may leave residual information in distributed representations (Section 6.2). The sharpest case is Wu et al. (2025)'s, since it holds where the proxy is exact: a model behaviourally indistinguishable from a retrained one still yields the forget data to an adversary reading it beside the pre-unlearning

version, and that access can be two logits APIs rather than two sets of weights. No test certifies, under adversarial conditions, that parameters no longer encode a given fact. Whether one can exist for models of a given class is not known, and a proof that it cannot would settle as much as the test would.

8.3 Theoretical Foundations

Nor does any theory say what would settle the distinction. Three fragments exist and none reaches the others: a divergence-rate separation between negative preference optimization and gradient ascent proved for logistic regression rather than a language model (Zhang et al., 2024), a gradient-alignment diagnostic scoring an objective's effect on forgetting against its effect on retention (Liu et al., 2025a), and a mechanistic account of which components store a fact rather than express it (Guo et al., 2024). A framework joining them would also have to range over released versions, since indistinguishability from the retrained oracle is necessary and not sufficient. Differential privacy, information-theoretic secrecy and cryptographic model release are where that connection lies.

8.4 Adversarial Robustness and Relearning Resistance

Meanwhile the knowledge returns through channels each defended one at a time. Robustness has to hold at two levels: the query, where a reformulated or optimized probe recovers content, and the weights, where a benign fine-tune or a routine quantization undoes the update. The defenses of Section 6.4 distribute across them one per channel: LAU against an optimized adversary, RULE against a sampling budget, ILU against a downstream fine-tune, MUDMAN against relearning. Each is validated against the channel it simulated, so the set is a coverage map with visible holes. Nothing here answers an inferential path through a knowledge graph or a second checkpoint, and the one claimed quantization defense comes from outside the family (Mao et al., 2025).

Each answers the threat it named, which leaves three questions none of them asks. Whether a defense holds against a threat it did not anticipate is untested. The strongest carry a cost (an adversarial inner loop, a sampling budget, a meta-learning stage) payable at a scale none has reached. And deployment applies the channels together rather than singly, while no standardized suite runs a model through fine-tuning, quantization and distillation in sequence.

8.5 Query-Invariant Unlearning and Response Shaping

The same open-endedness afflicts the query side, where unlearning is itself an attack surface. Gao et al. (2025) take the deletion request as an input the adversary controls: rewriting the submitted forget corpus through templates carrying an ordinary token (*please*, *then*) at a 2% insertion rate turns that token into a trigger, so later users who write it get refusals and hallucinations on retained knowledge, while forgetting is untouched. On TOFU with GradDiff on LLaMA at a 10% forget set the score reads a near-perfect 0.008, against 0.005 unattacked, while utility on the retained split falls from 0.683 to 0.160. The instrument reports success throughout. Their scope term recovers most of that loss; the remedies aimed at query-dependence itself cost more: a reasoning trace per forget item for TRU, and for Attention Shifting a suppression threshold trading unlearning, utility and hallucination against each other (Sections 6.2 and 6.4). None supplies query-invariance by design, and Jin et al. (2024)'s four probe types measure what it would have to survive.

8.6 Continual and Sequential Unlearning

And every one of these compounds when deletion requests keep arriving, which under a right-to-erasure regime is the ordinary case and not the stress case. Run in sequence on MUSE News, both negative preference optimization and SimNPO lose retained knowledge with each successive request; SimNPO loses it more slowly (Zhao et al., 2025). That is a better slope, not a flat one. Bae et al. (2026) avoid the drift by not touching the weights and hold flat across ten rounds. They then say in their own Limitations that deployment would bring thousands or millions of requests, and that bypassing the embedder routing them is untested.

8.7 Scalability to Frontier Models

All of it is demonstrated an order of magnitude below deployment scale. The methods reviewed here are unlearned at 7B to 13B, with two exceptions, both Li et al. (2024)'s: RMU is applied to a 34B model and to an 8×7B mixture of experts, and the authors report running no baseline at either size, so the corpus's two largest results have nothing to compare against. ECO and CURaTE sidestep the question by leaving the weights alone (Section 6.3). Whether weight modification stays effective at 70B and beyond is unanswered for the dominant paradigm.

8.8 From Certification to Comparative Measurement

Taken together these say the verification question is posed wrongly. We propose that it be answered comparatively rather than by certificate: for each method, how much of the target it hides, at what cost to retained capability, against a named adversary whose access is declared (inference only, the weights, or a second released version). Liu et al. (2025b) already ask for a stakeholder-indexed reporting card, and Shi et al. (2024) come nearest to supplying one, separating data-owner from deployer expectations. What neither supplies is the declared adversary that makes two such reports commensurable; the card is offered as a route to certification, and MUSE closes on a readiness verdict. Liu et al. (2024)'s *weak unlearning* shows it can be stated. Removal remains the aspiration. This is a claim about what today's evidence can support, not about what the field should stop wanting.

9 Conclusion

This survey has reviewed papers published between 2023 and 2026: five evaluation frameworks and platforms (Maini et al., 2024; Shi et al., 2024; Li et al., 2024; Jin et al., 2024; Kurmanji et al., 2025), twenty-one method and analysis papers, among them gradient-level and mechanistic accounts of what the objectives do to a model, and four studies of what the evaluations do not test (Gao et al., 2025; Hu et al., 2025; Zhang et al., 2025; Wu et al., 2025). The methods fall into four paradigms: output-space objectives that lower the probability of the target response (Yao et al., 2024; Eldan & Russinovich, 2023; Zhang et al., 2024; Zhao et al., 2025; Li et al., 2026), representation-space interventions that redirect or suppress the activations carrying it (Li et al., 2024; Mao et al., 2025; Ashuach et al., 2026; Tan et al., 2026), weight-preserving methods that intervene at inference and leave θ untouched (Liu et al., 2024; Ji et al., 2024; Bae et al., 2026), and robustness-by-design algorithms that anticipate a named attack (Yuan et al., 2024; Reisizadeh et al., 2026; Wang et al., 2025; Sondej et al., 2025; Liao et al., 2026).

Across that corpus the findings converge. Unlearning is cheap relative to retraining and no method satisfies every criterion at once; the forget–retain trade-off is why, rooted in the entanglement of knowledge across shared parameters. The instruments are fragmented across benchmarks and unstable within them. The sharpest convergence is on recovery. Knowledge certified as forgotten returns through sampling the model more than once, rephrasing the query, following inferential paths through a knowledge graph, appending an optimized adversarial suffix, fine-tuning on public text that never contained it, quantizing to four bits, and comparing a released model against its predecessor. Each of those seven channels can falsify a removal claim, and none can verify one. On that evidence these methods suppress the expression of knowledge rather than removing it from the weights.

What they converge on is a measurement problem rather than an optimization one. Of the seven problems Section 8 sets out, the load-bearing ones are about evidence: no test separates genuine forgetting from obfuscation, no theory says what would settle the question, and no benchmark reaches the scales or pipeline stages at which models are served. The proposal it closes on is comparative rather than certifying. Privacy law, copyright and safety each ask these methods for a guarantee, and a comparison is what the field can currently supply.

Acknowledgments

Acknowledge any assistance, funding, or resources used, if applicable.

References

Tomer Ashuach, Dana Arad, Aaron Mueller, Martin Tutek, and Yonatan Belinkov. CRISP: Persistent concept unlearning via sparse autoencoders. *arXiv preprint arXiv:2508.13650*, 2026.

Seyun Bae, Seokhan Lee, and Eunho Yang. CURaTE: Continual unlearning in real time with ensured preservation of LLM knowledge. *arXiv preprint arXiv:2604.14644*, 2026.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy*, pp. 141–159, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1877–1901, 2020.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy*, pp. 463–480, 2015.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Ronen Eldan and Mark Russinovich. Who's Harry Potter? Approximate unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), 2016.

Xiaodong Gao, Yixin Dong, Hao Li, and Hengrui Zhao. Keeping an eye on LLM unlearning: The hidden risk and remedy. *arXiv preprint arXiv:2506.00359*, 2025.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. *arXiv preprint arXiv:2410.12949*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *International Conference on Learning Representations (ICLR)*, 2025.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient LLM unlearning framework from logit difference. *arXiv preprint arXiv:2406.08607*, 2024.

Jinghan Jia, Jiancheng Liu, Yihua Mu, Parikshit Ram, Yuguang Bai, Aditi Raghunathan, and Sijia Liu. WAGLE: Strategic weight attribution for effective and modular unlearning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zhuoran Jin, Zhangdie Cao, Hongbang Chen, Yubo Zhang, Kang Liu, and Jun Zhao. RWKU: Benchmarking real-world knowledge unlearning for large language models. *arXiv preprint arXiv:2406.10890*, 2024.

Murat Kurmanji, Eleni Triantafillou, Jamie Hayes, and Peter Triantafillou. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

Yichong Leng, Shengqi Zhu, and Jingbo Shang. A closer look at machine unlearning for large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Kemou Li, Qizhou Wang, Yue Wang, Fengpeng Li, Jun Liu, Bo Han, and Jiantao Zhou. LLM unlearning with LLM beliefs. In *International Conference on Learning Representations (ICLR)*, 2026.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning (ICML)*, 2024.

Qing Li, Jiahui Geng, Herbert Woisetschläger, Zongxiong Chen, Fengyu Cai, Yuxia Wang, Preslav Nakov, Hans Arno Jacobsen, and Fakhri Karray. A survey of machine unlearning in large language models: Methods, challenges and future directions. *arXiv preprint arXiv:2503.01854*, 2025.

Junfeng Liao, Qizhou Wang, Shanshan Ye, Xin Yu, Ling Chen, and Zhen Fang. Explainable LLM unlearning through reasoning. In *International Conference on Learning Representations (ICLR)*, 2026.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jiancheng Liu, Jinghan Jia, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. Understanding the G-effect: How gradient-based unlearning objectives implicitly regulate model integrity in LLM unlearning. In *International Conference on Learning Representations (ICLR)*, 2025a.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 2025b.

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.

Ruiyang Mao, Shuang Wang, Zonghan Wang, Xuejing Liu, and Yao Ding. LLM unlearning via neural activation redirection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

OpenAI. GPT-4 technical report. *CoRR, abs/2303.08774*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hadi Reisizadeh, Jiajun Ruan, Yiwei Chen, Soumyadeep Pal, Sijia Liu, and Mingyi Hong. Leak@k: Unlearning does not make LLMs forget under probabilistic decoding. *arXiv preprint arXiv:2511.04934*, 2026.

Jiacheng Shi, Xiao Chen, Jiarun Hu, Hao Fei, Seyed Mehran Kazemi, and Huan He. MUSE: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pp. 3–18, 2017.

Filip Sondej, Yushi Yang, Mikołaj Kniejski, and Marcel Windys. Robust LLM unlearning with MUDMAN: Meta-unlearning with disruption masking and normalization. In *NeurIPS Workshop on Continual and Compatible Foundation Model Updates (CCFM)*, 2025.

Chenchen Tan, Youyang Qu, Xinghao Li, Hui Zhang, Shujie Cui, Cunjian Chen, and Longxiang Gao. Wisdom is knowing what not to say: Hallucination-free LLMs unlearning via attention shifting. *arXiv preprint arXiv:2510.17210*, 2026.

Changsheng Wang, Yihua Zhang, Jinghan Jia, Parikshit Ram, Dennis Wei, Yuguang Yao, Soumyadeep Pal, Nathalie Baracaldo, and Sijia Liu. Invariance makes LLM unlearning resilient even to unanticipated downstream fine-tuning. In *International Conference on Machine Learning (ICML)*, 2025.

Rongzhe Wei, Peizhi Niu, Hans Hao-Hsun Hsu, Ruihan Wu, Haoteng Yin, Mohsen Ghassemi, Yifan Li, Vamsi K. Potluru, Eli Chien, Kamalika Chaudhuri, Olgica Milenkovic, and Pan Li. Do LLMs really forget? evaluating unlearning with knowledge correlation and confidence awareness. *arXiv preprint arXiv:2506.05735*, 2025.

Xiaoyu Wu, Yifei Pang, Terrance Liu, and Zhiwei Steven Wu. Unlearned but not forgotten: Data extraction after exact unlearning in LLM. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hongbang Yuan, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. *arXiv preprint arXiv:2408.10682*, 2024.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

Zhiwei Zhang, Jiale Yue, Chenliang Li, Jinghan Jia, and Sijia Liu. Catastrophic failure of LLM unlearning via quantization. In *International Conference on Learning Representations (ICLR)*, 2025.

Chongyang Zhao, Bruno Lepri, Nicu Sebe, and Lixin Fan. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A Additional Details

A.1 RWKU Adversarial Probe Taxonomy

RWKU (Jin et al., 2024) evaluates forget efficacy under nine adversarial attack probe types, each designed to reformulate standard queries about the unlearning target through a different strategy. The full taxonomy is:

1. **Prefix Injection.** A contextual prefix that indirectly identifies the target is prepended to the query, e.g., "I've noticed there's a common author between the films *Misery* and *The Shining*. Could you confirm who wrote these?"
2. **Affirmative Suffix.** An affirmative suffix is appended to encourage the model to respond positively rather than refuse.
3. **Role Playing.** The model is assigned a role (e.g., "You are an expert biographer") and asked questions in that context.

4. **Multiple Choice.** The target knowledge is embedded in a multiple-choice format, testing whether the model can still select the correct option even when it declines to generate the answer directly.
5. **Reverse Query.** The question is posed from the opposite direction (e.g., asking which person matches a description rather than asking for a description of the person).
6. **Synonym Manipulation.** Synonyms or alternative phrasings of the target name or key terms are used to bypass pattern-level suppression.
7. **Background Hint.** Contextual clues about the target are provided before the question is asked, priming the model's recall.
8. **In-Context Learning.** Few-shot examples pointing toward the target are provided, leveraging the model's in-context learning ability to override the unlearning effect.
9. **Cross-Lingual.** The question is posed in a language other than English, testing whether unlearning generalizes across the model's multilingual capabilities.

Of the nine, the paper singles out four: "prefix injection, affirmative suffix, multiple choice and reverse query attacks effectively elicit unlearned knowledge from the model." Exposure is uneven across methods as well as across probes. Refusal tuning, being fine-tuned on refusal data, achieves the best unlearning efficacy under adversarial attack, and negative preference optimization also shows some resistance (Jin et al., 2024).

A.2 OpenUnlearning: Evaluation Inconsistency Details

OpenUnlearning (Kurmanji et al., 2025) provides an open-source modular framework implementing multiple unlearning methods and evaluation pipelines under a unified codebase. Its primary empirical contribution to the evaluation literature is a systematic demonstration that experimental setup choices significantly affect method rankings.

Key specific findings include:

- Hyperparameter sensitivity varies across methods: methods that appear robust to hyperparameter choices under one benchmark can be highly sensitive under another.
- Learning rate and number of gradient steps are the most consequential choices; small changes can flip the ranking between competing methods.
- Methods that perform well on TOFU do not necessarily perform well on MUSE under identical hyperparameters, reinforcing the cross-benchmark generalization gap discussed in the main text.
- Several published comparisons use non-identical experimental setups (different learning rates, different numbers of epochs, different retain-set compositions), making direct comparison unreliable.

The framework is publicly available and supports standardized evaluation across TOFU, MUSE, and additional benchmarks, providing a practical tool for future work on evaluation consistency.

A.3 LLM-as-a-Judge: Dimensions, Instantiation, and Scope Study

The LaaJ protocol of Liao et al. (2026) uses a reasoning judge model to score generations on six dimensions. Unlearning quality comprises relevance, rejection, and helpfulness; retention quality comprises readability, specificity, and logic. Each benchmark instantiates the dimensions against its own targets: on the fictitious-author benchmark the unlearning dimensions are scored on the real-authors subset and the retention dimensions on world facts; on the safety benchmark unlearning is scored on the hazardous question-answering subsets and retention on a general knowledge benchmark; and on the corpus benchmarks unlearning is scored on the verbatim- and knowledge-memorization forget sets and retention on the retain set. The reasoning model that generates the unlearning targets also serves as the judge, and the paper analyzes the relationship

between the two roles to check for bias. A separate scope study deliberately widens the unlearning scope from a narrow target to a broader category and shows that an imprecise scope induces incorrect refusals on queries that merely hint at the target, whereas a precisely specified scope preserves correct answers, illustrating that scope specification is itself a control the judge protocol can measure. The judge prompt templates and full per-dimension rubrics are provided by Liao et al. (2026).

A.4 Auxiliary Datasets Used by Method Papers

Several method papers evaluate on datasets outside the four benchmark frameworks. Sports Facts pairs athletes with the sport they play and supports both editing and unlearning of factual associations, while CounterFact provides subject-relation-object facts with alternative targets for counterfactual editing; both are used to study mechanistic localization (Guo et al., 2024). Pile-Bio is a molecular-biology subset of a large pretraining corpus used as a proxy for hazardous biological knowledge, with the remainder of the corpus serving as the retain set; CodeSearchNet function bodies are used to unlearn a coding skill while retaining general language ability; and a harmful-behavior dataset, BeaverTails, supplies prompts used to attempt the removal of a cruel tendency (Sondej et al., 2025). In each case the dataset is chosen to isolate a specific phenomenon, whether mechanistic localization, skill removal, tendency removal, or relearning robustness, rather than to serve as a general-purpose unlearning benchmark.

A.5 Benchmark Construction Sizes

The evaluation benchmarks of Section 5.1 differ in scale as well as in design. TOFU (Maini et al., 2024) comprises 200 fictitious author profiles, each described by 20 LLM-generated question–answer pairs, for 4,000 QA pairs in total, on which two base models are fine-tuned until the fictitious facts are memorized. MUSE (Shi et al., 2024) provides two real-text corpora: the NEWS corpus uses approximately 1.5M tokens of news articles as the forget set and 1.6M tokens of disjoint articles as the retain set, and the BOOKS corpus uses approximately 1.1M tokens from the Harry Potter series as the forget set and 0.5M tokens of Harry Potter FanWiki articles as the retain set. WMDP (Li et al., 2024) consists of 3,668 expert-authored multiple-choice questions distributed across three hazardous domains: WMDP-Bio (∼1,520 questions), WMDP-Cyber (∼1,987), and WMDP-Chem (∼408).

A.6 Surface- and Knowledge-Level Metric Definitions

The surface- and knowledge-level metrics of Section 5.2 are defined through token-overlap scores. MUSE's *verbatim memorization* (C1) prompts the model with the first l tokens of each forget-set sequence and compares the generated continuation to the true text via ROUGE-L:

$$\mathrm{VerbMem}(f, \mathcal{D}_f) = \frac{1}{|\mathcal{D}_f|} \sum_{x \in \mathcal{D}_f} \text{ROUGE-L}\big(f(x_{[:l]}),\ x_{[l+1:]}\big), \tag{14}$$

where f denotes the model's generation function and lower values indicate less verbatim reproduction. MUSE's *knowledge memorization* (C2) uses QA pairs (q, a) derived from the forget corpus via GPT-4 and averages the ROUGE overlap between the generated and reference answers:

$$\mathrm{KnowMem}(f, \mathcal{D}_f) = \frac{1}{|\mathcal{D}_f|} \sum_{(q,a) \in \mathcal{D}_f} \text{ROUGE}\big(f(q),\ a\big). \tag{15}$$

A.7 ECO Zeroth-Order Optimization Procedure

ECO (Liu et al., 2024) optimizes the embedding corruption magnitude σ using a zeroth-order optimization procedure that does not require gradient access to the model. The procedure:

1. Samples a batch of forget-set queries and retain-set queries.

2. For each candidate σ value, corrupts the forget-set embeddings with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and evaluates: (a) the model's loss on forget-set queries (should be high) and (b) the model's loss on retain-set queries (should remain low).

3. Uses finite-difference approximation of the gradient of a combined objective that maximizes forget-set loss while minimizing retain-set degradation.

4. Updates σ using the estimated gradient.

Because this procedure requires only forward passes through the model (no backpropagation), ECO can be applied to any model accessible through an inference API, including closed-weight models where only token-level output is available. The optimization converges in a few hundred evaluation steps and adds negligible overhead compared to weight-based unlearning methods.

A.8 Quantization Robustness: Weight-Change Magnitude Analysis

Zhang et al. (2025) show that existing unlearning methods produce weight perturbations $\Delta W = W_u - W_o$ whose magnitude is typically smaller than the quantization step size. For 4-bit quantization, the average step size is approximately 8×10^{-4}, while gradient ascent produces an average absolute weight change of approximately 3×10^{-4}. When $|\Delta w| < \Delta/2$ (half the quantization step), rounding maps the unlearned weight back to the same quantized value as the original weight, effectively reversing the unlearning. The paper reports that roughly 80% of weight changes fall below this threshold under standard unlearning hyperparameters.

SURE (Saliency-Based Unlearning with a Large Learning Rate) mitigates this by combining two components: (1) a deliberately enlarged learning rate that produces larger weight perturbations surviving quantization, and (2) a saliency mask (based on forget-loss gradient magnitudes, thresholded at the 90th percentile) that confines the aggressive updates to the parameters most relevant to the forget set, preventing utility degradation. SURE is a wrapper applicable to any fine-tuning-based unlearning method and does not define its own unlearning objective.

A.9 ULD: Assistant Construction and Decoding Details

The assistant model in ULD (Ji et al., 2024) is built from the target model itself, reusing its first K transformer layers together with the language-model head, so that the assistant shares the target's vocabulary and can be combined with it through logit arithmetic. Only lightweight low-rank adapters attached to these inherited layers are trained, while the inherited weights stay frozen. This keeps the trainable parameter count at a small fraction of the target model and is the main source of ULD's training efficiency. The forget objective is a cross-entropy loss on the forget documents, augmented with paraphrases so that the assistant generalizes across surface forms, while the retain objective drives the assistant's distribution on retain documents toward uniform, optionally augmented with perturbed answers that encode incorrect variants of the forget knowledge. At decode time the assistant logits are subtracted from the target logits with a scaling weight that controls forgetting strength, and a logit-filter step restricts the subtraction to tokens the target assigns non-negligible probability, preventing low-probability noise from distorting the output. The augmentation prompts and the specific choices of K, adapter rank, scaling weight, and filter threshold are reported by Ji et al. (2024).

A.10 LAU: Perturbation Placement and Attack-Stage Ablations

The latent perturbation in LAU (Yuan et al., 2024) is added to the residual stream at a chosen transformer layer, and its placement matters. Perturbing shallow layers, closer to the input, has the largest effect on the output and is easiest to optimize, whereas perturbing progressively deeper layers reduces the perturbation's influence until the method approaches its non-adversarial base objective. Perturbing directly at the embedding layer is ineffective, because the latent perturbation is meant to approximate the effect of an adversarial query rather than to edit the prompt. The number of inner optimization steps that build

the perturbation also trades off: too few or too many both weaken unlearning, though even a small, randomly initialized perturbation improves robustness over the non-adversarial baseline. Beyond the real-world knowledge benchmark reported in the main text, Yuan et al. (2024) confirm that the LAU-augmented variants remain effective on the book- and news-corpus setting, and that models trained with LAU resist the companion dynamic attack more strongly than models trained without it.

A.11 MUDMAN: Module Selection, Normalization, and Rejected Components

Sondej et al. (2025) find that the choice of which modules to update strongly affects the separation between forgetting and retention. The first feed-forward projection of each block, the layer whose output passes through the activation nonlinearity, gives the cleanest separation, since updates there can deactivate neurons in a way that resists relearning, whereas the attention query and key projections and the second feed-forward projection tend to disrupt retained performance without contributing much to forgetting. Freezing the unhelpful modules also saves memory. Among normalization strategies, a global norm computed across the intervened modules works best, and normalizing before rather than after the disruption mask is chosen for simplicity. The paper also documents an extensive catalogue of components that did not help, including several schemes for dampening relearning gradients, direct weight ablation by neuron or by weight relevance, restricting updates to shrink weights or activations, a selective logit loss, multiple adversaries, an unlearning gradient accumulator, and variants that keep the adversary synchronized with the main model; most were outperformed by the simpler combination adopted in the final method. A compact implementation of the single-loop procedure is provided by Sondej et al. (2025).

A.12 ILU: Invariance Formulation and Baseline Comparisons

ILU (Wang et al., 2025) derives its penalty from the practical form of invariant risk minimization, in which the bilevel invariance objective is relaxed to a single-level problem by penalizing the squared gradient of the per-environment loss with respect to a fixed scalar predictor evaluated at unity. Treating each fine-tuning task as an environment, this penalty is added to the unlearning objective, so that the unlearned solution is stationary under fine-tuning perturbations:

$$\min_{\theta} \; \mathcal{L}_{\mathrm{u}}(\theta) \; + \; \lambda \left\| \nabla_{w} \, \mathcal{L}_{\mathrm{ft}}(w \cdot \theta;\, \mathcal{D}) \big|_{w=1} \right\|_{2}^{2} \tag{16}$$

where $\mathcal{L}_{\mathrm{u}}$ is the base unlearning objective, $\mathcal{L}_{\mathrm{ft}}$ the loss on the unlearning-irrelevant fine-tuning set $\mathcal{D}$, and w the fixed scalar predictor. The authors show that using a single unlearning-irrelevant fine-tuning set for the penalty generalizes better than using the forget set itself or a mixture of several fine-tuning sets, the latter introducing optimization conflicts that can destabilize utility. Compared against other robustness-oriented approaches, latent adversarial training yields only marginal gains over the non-robust baseline, while a meta-learning tamper-resistance method reaches comparable robustness to ILU but at far higher computational cost; ILU attains similar robustness with a large efficiency advantage. The approach also transfers to the book- and news-corpus benchmark, where it keeps verbatim- and knowledge-memorization scores low after downstream fine-tuning that otherwise restores them.

A.13 FLU: Localization Methodology and Parameter-Count Controls

The fact-lookup localization in Guo et al. (2024) is derived differently for the two datasets studied. For the sports-facts setting, it follows prior mechanistic analysis and uses linear probes on intermediate activations to identify the layers at which a subject's representation becomes predictive of the target attribute, taking the feed-forward components at those layers as the localization. For the counterfactual setting, where answers do not fall into a small set of classes, it uses path patching, first to isolate the components that write the fact to the output and then to find the feed-forward components that feed those extraction components, which are taken as the fact-lookup mechanism. To ensure that comparisons are not confounded by the size of the edit, the authors control the number of updated parameters across localizations and additionally train binary weight masks to equalize the exact parameter count; the robustness advantage of fact-lookup localization persists under these controls. Analysis of which mechanism each localization actually modifies

confirms that output-tracing and non-localized editing change the extraction components more than the fact-lookup components, consistent with their weaker robustness. Per-model probe curves and the full set of localization comparisons are reported by Guo et al. (2024).

A.14 RMU: Layer Selection, Coreset, and Compute

Representation Misdirection for Unlearning (RMU) (Li et al., 2024) is applied at an intermediate transformer layer, typically in the range 5–10 for 7B-parameter models. Ablations show that applying it too early (layers 1–3) is ineffective, because later layers can reconstruct the corrupted representations, while applying it too late (layers 28–32) damages general capabilities, because high-level features are shared across domains; middle layers (5–10), where domain-specific representations are formed, give the best forget–retain trade-off. The coreset-selection procedure identifies a small subset (∼500–1,000 samples) of the forget set that maximally reduces WMDP accuracy per gradient step, making the procedure practical even with large forget corpora. The full procedure requires fewer than 500 gradient steps and under one GPU-hour for 7B-parameter models.

A.15 Bootstrapping Framework: BS-T and BS-S Objectives

Li et al. (2026) derive the token-level (BS-T) and sequence-level (BS-S) objectives to counter probability mass redistribution (the squeezing effect). In BS-T, given an input x and the current model's output distribution $P_\theta(\cdot \mid x)$, the loss incorporates the soft prediction vector across the vocabulary $\mathcal{V}$ rather than a one-hot target:

$$\mathcal{L}_{\text{BS-T}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}_f}\left[\sum_{v\in\mathcal{V}} P_{\text{orig}}(v \mid x)\log\left(1 - P_\theta(v \mid x)\right)\right] \tag{17}$$

where P_{orig} provides reference soft targets. This directly penalizes tokens assigned high probability by the original model, attenuating probability mass in squeezed regions. In BS-S, for each prompt $x \in \mathcal{D}_f$, N high-confidence completions $\mathcal{A}_{\text{seq}} = \{y_1, y_2, \ldots, y_N\}$ are sampled from the original model $P_{\text{orig}}(\cdot \mid x)$ using temperature sampling. The unlearn dataset is augmented with these sequence beliefs, and the unlearning objective suppresses them jointly:

$$\mathcal{L}_{\text{BS-S}}(\theta) = \mathcal{L}_{\text{base}}(\theta; x, y_{\text{target}}) + \frac{\lambda}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{base}}(\theta; x, y_i) \tag{18}$$

where $\mathcal{L}_{\text{base}}$ can be instantiated with NPO or GA. Experiments confirm that BS-T and BS-S eliminate high-confidence rephrasings and spurious unlearning across TOFU and MUSE.

A.16 CRISP: SAE Feature Selection and Fine-Tuning Details

CRISP (Ashuach et al., 2026) utilizes pre-trained Sparse Autoencoders (SAEs) from Llama Scope and Gemma Scope across intermediate transformer layers. The target feature selection algorithm computes activation differentials for each SAE feature $f_i \in \mathcal{F}_l$ at layer l:

$$\Delta\text{Act}(f_i) = \mathbb{E}_{x\in\mathcal{D}_{\text{target}}}[f_i(\mathbf{h}_l(x))] - \mathbb{E}_{x\in\mathcal{D}_{\text{retain}}}[f_i(\mathbf{h}_l(x))] \tag{19}$$

Features satisfying $\Delta\text{Act}(f_i) > \tau$ and demonstrating high monosemantic alignment with hazardous domain concepts (e.g., viral infection or network exploit terms) are selected as target set $\mathcal{F}_{\text{target}}$. During fine-tuning, parameter updates minimize:

$$\mathcal{L}_{\text{CRISP}}(\theta) = \sum_{f\in\mathcal{F}_{\text{target}}}\mathbb{E}_{x\in\mathcal{D}_{\text{target}}}[f(\mathbf{h}_l(x;\theta))] + \gamma\sum_{f\in\mathcal{F}_{\text{benign}}}\mathbb{E}_{x\in\mathcal{D}_{\text{retain}}}\|f(\mathbf{h}_l(x;\theta)) - f(\mathbf{h}_l^{\text{orig}}(x))\|^2 \tag{20}$$

Optimization requires fewer than 200 gradient steps, producing permanent weight modifications while preserving benign feature activations.

A.17 Attention Shifting: Entropy-Based Token Importance and Dual-Loss Adapters

Tan et al. (2026) formulate token importance $I(t_i)$ for an input sequence $x = \{t_1, t_2, \ldots, t_n\}$ via predictive entropy changes under masking:

$$I(t_i) = H(P_\theta(\cdot \mid x)) - H(P_\theta(\cdot \mid x_{-i})) \tag{21}$$

where $H(P) = -\sum_v P(v) \log P(v)$ is predictive entropy and x_{-i} is the sequence with token t_i masked. Higher $I(t_i)$ values denote factual anchors (proper nouns, domain terms). Importance scores are then used to build a reweighted, renormalized target attention distribution, and the objective is a convex combination $\alpha\mathcal{L}_{\text{ASP}} + (1-\alpha)\mathcal{L}_{\text{AKL}}$ of two KL divergences between attention distributions. The two are taken in opposite directions: on the forget set from the model's attention to the target, and on the retain set from the target back to the model. Fine-tuning is restricted to low-rank adapter (LoRA) weights injected into key/value attention projections.

A.18 CURaTE: Real-Time Embedding Retrieval and Hard Negative Training

CURaTE (Bae et al., 2026) trains an unlearning sentence embedder $E_\phi(\cdot)$ prior to LLM deployment using the classical margin-based contrastive loss of Hadsell et al., which the paper cites by name: cosine distance pulls each forget query q_i toward a semantically equivalent paraphrase q_i^+ and pushes it away from hard negatives $q_{i,j}^-$ until they exceed a margin m. The hard negatives are queries sharing entity names or topical keywords but requesting benign information. Post-deployment, forget request embeddings $\{E_\phi(x_f)\}_{x_f \in \mathcal{D}_f}$ are stored in a FAISS vector index. For an incoming query x_{user}, the maximum similarity $S = \max_{x_f} \text{sim}(E_\phi(x_{\text{user}}), E_\phi(x_f))$ is computed. If $S \geq \theta_{\text{refusal}}$, a pre-formatted refusal message is returned immediately; otherwise, x_{user} is routed to the target LLM. Real-time insertion latency is sub-millisecond per request.

A.19 RULE: Iterative Self-Generation and Forget-Set Augmentation

Reisizadeh et al. (2026) propose Robust Unlearning under Leak@k (RULE) to defend against multi-sample stochastic generation leakage. RULE changes no loss function. Given a prompt q from the forget set, it samples k generations from the current model under probabilistic decoding, scores them, and admits those clearing a threshold τ into an augmented forget set; the base objective (GradDiff in the reported experiments) is then re-run on the augmented data. The cycle repeats for three iterations, so the leakage that only stochastic decoding exposes becomes training data for the next round.

The decoding sweep varies temperature and top-p jointly, with top-p the primary driver of leakage. Suppression across large sampling budgets ($k \geq 64$) holds on TOFU; it does not hold on MUSE.

A.20 Benign Relearning: Relearn Set Construction

Hu et al. (2025) construct benchmark-specific relearn sets designed to be topically related to the forget set but uninformative about its specific content:

- **TOFU.** For each fictitious author, a single book title is provided as a keyword prompt, and the model generates text seeded by this keyword. No biographical facts (nationality, birth year, writing style) appear in the relearn data.
- **WHP (Who's Harry Potter).** The relearn set is GPT-4-generated general information about the main characters, who are named in it. It "does not contain direct text from the novels," so the memorized source prose the unlearning targets is absent.
- **WMDP.** General-domain scientific articles in the relevant fields (biology, cybersecurity) are collected from public sources, excluding any content that overlaps with the hazardous knowledge targeted by the benchmark.

In each case, the authors verify that a model fine-tuned *only* on the relearn data (without prior exposure to the forget set) cannot answer the evaluation queries, confirming that the relearn data is uninformative about the targeted content.

A.21 Guidance-Based Extraction: Token Filtering Procedure

The guidance-based extraction method of Wu et al. (2025) augments its logit steering with a token filtering step. At each generation step, the set of candidate next tokens is restricted to those whose probability under the pre-unlearning model exceeds a threshold. This eliminates tokens that are unlikely under the original model, which contribute noise to the guided generation and reduce extraction precision. The combination of logit guidance and token filtering consistently outperforms either technique in isolation, and the guidance scale w is treated as a hyperparameter tuned on a small held-out set. The method requires only forward-pass access to both models and adds negligible computational overhead beyond standard autoregressive generation.

A.22 Dynamic Unlearning Attack: Construction and Generalization

The Dynamic Unlearning Attack of Yuan et al. (2024) optimizes an adversarial suffix through a gradient-guided discrete search over tokens, in the style of automated jailbreak attacks: at each step the gradient identifies promising candidate tokens, which are then evaluated directly and the best retained, and a single suffix is optimized jointly across several prompts so that it transfers as a universal trigger. Formally, the suffix q is chosen to maximize the likelihood of the target response y when appended to a query x,

$$\min_{q} \; -\sum_{i} \log p\big(y_i \mid [\,x;\, q;\, y_{<i}\,]\big), \tag{22}$$

where $[\,\cdot\,]$ denotes concatenation; the objective is optimized over the discrete tokens of q alone, with the model left frozen. The attack is characterized along two axes. The practicality axis varies the model used to optimize the suffix: the ideal case optimizes against the unlearned model itself, while the more realistic case optimizes against the original model, to which an attacker is more likely to have access. The generalization axis varies how far the optimized suffix must travel from its training queries: a within-query setting reuses the suffix on the same questions, a cross-query setting applies it to new questions about the same target, and a cross-target setting applies it to questions about entirely different targets. Recovery is strongest in the within-query setting and weakens with distance, but it remains well above the no-attack baseline even when the suffix is optimized on the original model and transferred, which is the observation that removes the assumption of unlearned-model access. A static baseline, in which adversarial questions are generated by a capable language model rather than optimized, is consistently weaker than the dynamic attack.