Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Diagnosing Recoverability from Early Editing Trajectories in Large Language Model Knowledge Editing

**Qi SUN**
North China University of Technology
**Jianyong DUAN**
duanjy_NCUT@163.com

North China University of Technology
**Na YAN**
Sinopec Petroleum Engineering Technology Research Institute Co., LTD

**Article**

**Keywords:** Knowledge editing, Large language models, Recoverability diagnosis, Editing trajectories, Adaptive knowledge editing, Model interpretability

**Posted Date:** November 20th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11301643/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Diagnosing Recoverability from Early Editing Trajectories in Large Language Model Knowledge Editing

Qi SUN[1,2], Jianyong DUAN[1,*] ,Na YAN[2]

1. North China University of Technology, Beijing 100144, China

2. Sinopec Petroleum Engineering Technology Research Institute Co., LTD, Beijing, China

* Correspondence: duanjy@ncut.edu.cn

**Abstract**

Knowledge editing aims to modify specific factual associations in large language models without retraining the entire model. Existing evaluations typically assess editing outcomes after a fixed editing procedure, leaving open whether an intermediate failure could still recover if additional editing capacity were allocated. We study this problem by observing how individual editing requests evolve under progressively increasing editing capacity. Intermediate failures are categorized as RECOVERABLE if they subsequently achieve joint rewrite and generalization success, and RESISTANT if they remain unsuccessful after further capacity expansion. We identify an early trajectory signal that measures the improvement of the weakest paraphrase path before the diagnostic point. On a fresh 500-case CounterFact confirmation set with GPT-J-6B, the frozen primary signal achieves an ROC-AUC of 0.7340 for distinguishing recoverable from resistant failures. A diagnostic policy using a threshold determined on separate development data retains 85.71% of recoverable cases while reducing additional capacity evaluations by 38.75%. The same signal remains informative on ZsRE and GPT-2 XL, with ROC-AUC values of 0.9316 and 0.7965, respectively. An independent representation-level analysis shows stronger cross-prompt alignment of edit-induced representation shifts in recoverable than in resistant failures. Overall, intermediate editing failures exhibit distinct subsequent trajectories, and their early dynamics provide useful information for deciding whether further editing capacity is warranted.

**Keywords: Knowledge editing; Large language models; Recoverability diagnosis; Editing trajectories; Adaptive knowledge editing; Model interpretability**

## 1 Introduction

Large language models (LLMs) acquire substantial factual knowledge during pretraining, but the knowledge encoded in their parameters is neither perfectly accurate nor permanently up to date. As real-world facts evolve and erroneous associations are identified, it becomes increasingly important to update specific model knowledge without repeatedly retraining the entire model. Knowledge editing addresses this problem by modifying targeted factual associations while seeking to preserve unrelated model behavior. A growing body of work has demonstrated that individual or multiple factual associations can be efficiently updated through localized parameter modifications, external memory, or learned editing mechanisms [1–6].

Despite substantial progress, existing knowledge-editing research predominantly evaluates an edit from its final post-edit state. An editing configuration is typically applied with a predetermined update scope, after which success is assessed in terms of rewrite accuracy, paraphrase generalization, locality, or related metrics[5,6]. Under this evaluation paradigm, unsuccessful edits are generally treated as a single category of

failures. However, such a binary view overlooks the fact that an edit may evolve progressively as additional editing capacity is allocated. An intermediate failure may represent either an edit that is still progressing toward successful generalization or one that is unlikely to benefit from further intervention.

This observation motivates a different question: when an edit has not yet succeeded, can we determine whether it is still recoverable before allocating additional editing capacity? If recoverable and persistently resistant failures can be distinguished from their early behavior, editing capacity could be allocated selectively rather than uniformly. This shifts the focus from predicting whether an edit is currently successful to diagnosing whether an unsuccessful edit is likely to benefit from continued editing.

In this work, we formulate knowledge-editing recoverability diagnosis, which characterizes whether an intermediate editing failure can subsequently achieve successful rewriting and generalization when additional editing capacity is introduced. Using MEMIT as the underlying editor, we construct progressive editing trajectories and distinguish intermediate failures into RECOVERABLE and RESISTANT cases according to their subsequent response to increased editing capacity. We then investigate whether information available before the continuation decision is sufficient to distinguish these two outcomes. In particular, we track the same difficult paraphrase across early editing stages and use its recent change in target preference as a diagnostic signal of continued editing progress.

Experiments on CounterFact and ZsRE with GPT-J-6B and GPT-2 XL show that intermediate failures can follow different subsequent trajectories and that early editing dynamics contain information about this recoverability. A frozen diagnostic rule selectively reduces continued editing while retaining most recoverable cases. Recoverable and resistant failures also differ in cross-prompt representation-shift alignment, providing an internal correlate of the behavioral distinction. The diagnostic signal is observed across both datasets and model architectures, although its numerical stopping threshold may require distribution-specific calibration.

Our main contributions are summarized as follows:

- Recoverability formulation. We formulate recoverability as a property of intermediate editing failures, distinguishing cases that subsequently achieve joint success under additional editing capacity from those that remain resistant.
- Early-trajectory diagnosis and adaptive stopping. We identify an early behavioral trajectory signal associated with subsequent recoverability and use it in a frozen stopping rule to selectively allocate additional editing capacity. The signal is evaluated across two datasets and two language models.
- Representation-level evidence. We show that recoverable and resistant failures differ in cross-prompt alignment of edit-induced representation shifts, providing representation-level evidence associated with the behavioral distinction.

## 2 Related Work

### 2.1 Knowledge Editing in Large Language Models

Knowledge-editing methods can broadly be divided into learned or memory-based editors and direct parameter-modification approaches. KnowledgeEditor and MEND learn auxiliary mechanisms for efficient model updates [1,2], while SERAC stores edits externally and selectively routes relevant inputs to a counterfactual model [3]. Parameter-modification methods instead identify and modify internal components associated with factual knowledge. Knowledge Neurons analyzes factual representations at the neuron level [4], ROME performs rank-one updates to selected feed-forward layers [5], and MEMIT extends this approach to large-scale editing by distributing updates across multiple mediating layers [6].

Existing methods primarily evaluate the outcome after applying a given editing configuration. In contrast, we use progressive MEMIT editing to examine how individual edits evolve as editing capacity increases.

### 2.2 Reliability and Generalization of Model Editing

Beyond successful modification of the original editing request, reliable knowledge editing requires the updated association to generalize to semantically equivalent inputs while minimizing unintended changes to unrelated knowledge. Consequently, commonly used evaluation protocols measure multiple properties, including rewrite success, paraphrase generalization, locality or specificity, and fluency [5,6,8]. CounterFact and ZsRE have become widely used benchmarks for evaluating these dimensions under factual editing [5,7].

Subsequent studies have shown that high rewrite accuracy alone does not necessarily imply robust editing. Edited knowledge may fail to generalize across alternative formulations, interfere with related facts, or exhibit unexpected behavior under broader evaluation settings [9–15]. These observations have motivated increasingly comprehensive evaluations of editing reliability, including portability, locality, robustness, and downstream effects [9,10,12–15].

Rather than asking only whether an edit generalizes after editing is complete, we ask whether an edit that currently fails the joint success criterion is still progressing toward successful generalization. This distinction is central to recoverability: two cases may exhibit the same failure status at an intermediate checkpoint while responding differently to subsequent editing capacity.

### 2.3 Adaptive and Diagnostic Perspectives on Knowledge Editing

Recent studies have moved beyond uniform editing configurations toward case-specific adaptation. DR-IKE adjusts retrieved editing context according to task difficulty [16], while Knowledge Spectrum characterizes case-specific editing difficulty to guide intervention strength [18]. TangleScore similarly estimates the intrinsic difficulty of individual knowledge instances and adapts editing strength accordingly [19]. For parameter editing, CAKE introduces causal-guided adaptive layer selection [20], and FiDAL uses Fisher information for instance-dependent localization [21]. These methods adapt the editing procedure according to properties estimated for individual requests, but their primary objective is to determine how an edit should be performed rather than whether an already observed failure remains recoverable.

Other studies examine editing dynamics and post-edit reliability. Lifelong editing work considers stopping and regularization under accumulated sequential edits [17], TamEdit models trajectories across continual editing operations [22], and uncertainty-based analysis identifies hidden instabilities after editing [23]. These settings differ from the progressive trajectory considered here, where multiple intermediate states of the same editing request are observed as capacity increases.

We focus specifically on recoverability after an intermediate failure: whether a case that has failed at the diagnostic capacity can still achieve joint success under further capacity expansion. The diagnosis uses only the behavioral trajectory observed before the continuation decision. This differs from pre-edit difficulty estimation and adaptive localization, as well as from trajectory modeling across successive editing requests and post-edit uncertainty analysis.

## 3 Early-Trajectory Diagnosis of Knowledge-Editing Recoverability

3.1 Problem Formulation

Consider a knowledge-editing request consisting of a rewrite prompt associated with a target answer and a set of paraphrase prompts expressing the same underlying fact. Given a pretrained language model $f_\theta$, an editing method progressively allocates editing capacity to produce a sequence of edited models $f_\theta^{(1)}$, $f_\theta^{(2)}$, ..., $f_\theta^{(K)}$, where k denotes the amount of editing capacity allocated to the current case. In our experiments, capacity is instantiated by the number of MEMIT editing layers used for the update.

We evaluate each edited model using a joint success criterion that requires both successful rewriting of the target fact and successful generalization to its paraphrase prompts. Let $S_k \in \{0,1\}$ denote whether an editing case satisfies this criterion at capacity k. Rather than treating a failure observed at an intermediate capacity as a terminal outcome, we distinguish cases according to whether the failure can subsequently be resolved by allocating additional editing capacity.

For a diagnostic capacity $k_d$ and a maximum evaluated capacity K, an intermediate failure is defined as RECOVERABLE if it fails at $k_d$ but succeeds at some larger capacity, and RESISTANT if it remains unsuccessful throughout the remaining capacity range:

$$y = \begin{cases} \text{RECOVERABLE}, & S_{k_d} = 0 \text{ and } \exists k \in \{k_d + 1, \dots, K\}: S_k = 1, \\ \text{RESISTANT}, & S_k = 0, \quad \forall k \in \{k_d, \dots, K\}. \end{cases} \tag{1}$$

This formulation defines a forward-looking diagnostic problem: given only information observed up to $k_d$, determine whether an intermediate failure is likely to benefit from additional editing capacity. In the main experiments, we set $k_d$ =3; later capacities are used only to establish the ground-truth recoverability label.

## 3.2 Progressive Editing Trajectories

To characterize how an edit evolves before the diagnostic point, we construct a progressive editing trajectory by evaluating the same case after successive capacity levels. At each k, the edited model is evaluated on both the direct rewrite and its paraphrase prompts. This produces a sequence of intermediate behavioral states rather than a single post-edit measurement.

For each paraphrase prompt i, we quantify target preference using the probability difference between the desired target and the original answer:

$$d_i^{(k)} = P_{f_\theta^{(k)}}(o_{\text{new}} \mid x_i) - P_{f_\theta^{(k)}}(o_{\text{old}} \mid x_i) \tag{2}$$

where $o_{new}$ and $o_{old}$ denote the new and original objects, respectively. A larger value therefore indicates stronger preference for the edited association under the corresponding paraphrase.

Importantly, prompt identity is preserved across capacity levels when constructing a trajectory. Rather than selecting a different difficult paraphrase at each capacity, we identify a single paraphrase at the initial capacity and track that same prompt throughout subsequent editing stages. The exact selection rule is defined in Section

3.4. This design ensures that changes in the trajectory reflect the response of the same prompt to increasing editing capacity rather than changes in prompt identity.

## 3.3 Recoverable and Resistant Failures

Recoverability is evaluated only for cases that remain unsuccessful at the diagnostic capacity. A particularly relevant setting arises when the direct rewrite has already succeeded but generalization to semantically equivalent paraphrases remains incomplete. Additional capacity may allow the edited association to propagate to these paraphrases, or the case may remain unsuccessful despite further expansion. The RECOVERABLE/RESISTANT distinction captures these different responses to continued editing.

## 3.4 Early-Trajectory Diagnostic Signals

At $k = 1$, we identify the weakest failed paraphrase $i^*$ as the failed prompt with the smallest target-preference probability difference. If all paraphrases succeed at $k = 1$, we instead select the paraphrase with the smallest probability difference among all paraphrases. The identity of $i^*$ is then fixed and tracked at all subsequent capacity levels.

Let $d_{i^*}^{(k)}$ denote the probability difference between the edited and original targets for this tracked prompt at capacity $k$. Our primary trajectory feature is the adjacent gain from $k = 2$ to $k = 3$:

$$g_{2\to3} = d_{i^*}^{(3)} - d_{i^*}^{(2)} \tag{3}$$

A larger positive value indicates that the weakest early paraphrase path is still improving as editing capacity approaches the diagnostic point, whereas a small or negative value indicates early saturation or deterioration. The feature $g_{2\to3}$ was selected during development and frozen as the primary diagnostic signal before evaluation on the fresh confirmation set. Other trajectory summaries, including $g_{1\to2}$ and $g_{1\to3}$, are therefore treated as secondary analyses rather than alternative primary endpoints.

## 3.5 Diagnostic-Guided Stopping

The diagnostic signal can be converted into a simple stopping policy at the intermediate capacity $k = 3$. Given a threshold $\tau$, the decision rule is:

$$\pi(g_{2\to3}) = \begin{cases} \text{CONTINUE}, & g_{2\to3} \geq \tau, \\ \text{STOP}, & g_{2\to3} < \tau. \end{cases} \tag{4}$$

CONTINUE indicates that the trajectory provides sufficient evidence to allocate additional editing capacity, whereas STOP indicates that further expansion is not selected under the diagnostic rule. The threshold is estimated exclusively from development data and is fixed before confirmation evaluation. The purpose of this rule is not to maximize classification accuracy in isolation, but to reduce unnecessary editing while retaining as many genuinely recoverable cases as possible.

**4 Experimental Setup**

## 4.1 Models and Datasets

We evaluate the proposed recoverability-diagnosis framework using two autoregressive language models, GPT-J-6B[24] and GPT-2 XL[25], and two widely used knowledge-editing benchmarks, CounterFact and ZsRE. The main experiments are conducted on GPT-J-6B with CounterFact. To examine whether the observed diagnostic pattern generalizes beyond a single benchmark, we additionally evaluate GPT-J-6B on ZsRE. Cross-model validation is further performed on GPT-2 XL with CounterFact.

CounterFact [5] consists of factual editing requests designed to replace an existing factual association with a counterfactual target, together with paraphrase prompts that evaluate whether the edited association generalizes across alternative formulations. ZsRE [7] provides question–answer-based factual editing requests and corresponding rephrased queries, providing a complementary evaluation setting for assessing generalization after editing.

For the main CounterFact experiment, we use a fresh evaluation set of 500 cases that is separate from the cases used during earlier development and diagnostic exploration. The ZsRE cross-dataset validation also contains 500 cases. Cross-model validation on GPT-2 XL is conducted on 200 CounterFact cases. Recoverability analysis is performed only on cases that remain unsuccessful at the diagnostic capacity $\mathrm{k} = 3$; cases that have already achieved joint success by this point do not require a continuation decision.

## 4.2 Editing Configuration

We use MEMIT [6] as the underlying knowledge-editing method and construct progressive editing trajectories by varying the number of editing layers allocated to each request. In this work, editing capacity $k$ is therefore operationalized as the number of MEMIT candidate layers included in the update. For a given case, increasing $k$ progressively expands the editing configuration while leaving the underlying editing method unchanged. This design allows the response of the same editing request to be observed under increasing editing capacity. Accordingly, editing capacity should be interpreted here as a controlled, MEMIT-specific operational variable rather than a model-independent measure of editing strength.

For the main GPT-J-6B experiments, editing capacity is progressively increased from $k = 1$ to $k = 6$, with $k = 3$ designated as the diagnostic point. The diagnostic point is positioned at an intermediate capacity so that sufficient early trajectory information is available from $k = 1 - 3$ while additional capacity remains available for observing subsequent recovery at $k = 4 - 6$. Cases that have not achieved joint success at $k = 3$ are subsequently evaluated at $k = 4$, $k = 5$, and $k = 6$ to establish their recoverability labels. These later outcomes are not used to construct the diagnostic feature.

For GPT-J-6B, we use the six MEMIT candidate layers $\{3,4,5,6,7,8\}$, while for GPT-2 XL we use the five candidate layers $\{13,14,15,16,17\}$. Progressive editing starts from the last candidate layer and expands toward earlier layers: at capacity $k$, the last $k$ layers in the corresponding candidate-layer list are included in the MEMIT update. Thus, for example, $k = 3$ corresponds to layers $\{6,7,8\}$ for GPT-J-6B and $\{15,16,17\}$ for GPT-2 XL. All other MEMIT hyperparameters are held fixed across capacity levels.

We follow the corresponding MEMIT configurations with subject-last fact-token selection, a learning rate of 0.5, a clamp-norm factor of 0.75, and second-moment statistics estimated from 100,000 Wikipedia samples. The value-optimization procedure uses 25 gradient steps for GPT-J-6B and 20 steps for GPT-2 XL.

For the GPT-2 XL cross-model experiment, we use the standard five-layer MEMIT candidate configuration and progressively evaluate $k = 1, \ldots, 5$, again using $k = 3$ as the diagnostic point. Thus, the operational definition of recoverability and the diagnostic point remain unchanged across models, while the maximum available editing capacity follows the model-specific MEMIT configuration.

## 4.3 Evaluation Protocol

At each capacity level, editing success is evaluated on both the direct rewrite and its paraphrase/generalization prompts. For CounterFact, a prompt is considered successful when the edited target has lower negative log-likelihood (NLL) than the original target. An edit satisfies the joint success criterion only when the rewrite criterion is satisfied and all evaluated paraphrase prompts are successful. For ZsRE, rewrite and rephrased generalization are evaluated using teacher-forced target-token prediction. A prompt is considered successful only when the model's top-1 prediction matches the target token at every position in the target sequence. Each case contains one rephrased prompt, and joint success requires both the rewrite and rephrased prompt to satisfy this sequence-level criterion.

Each editing case is first evaluated progressively through the diagnostic capacity $k = 3$. The primary diagnostic feature $g_{2\to3}$ is computed entirely from observations available by this point. Cases satisfying the joint success criterion at $k = 3$ are excluded from the recoverability analysis. For the remaining intermediate failures, editing is continued to higher capacities only to establish the retrospective ground-truth label. A case is labeled RECOVERABLE if it subsequently achieves joint success at any evaluated capacity $k > 3$, and RESISTANT if it remains unsuccessful through the maximum evaluated capacity.

To avoid test-set tuning, the primary diagnostic feature definition and stopping rule were frozen before the fresh 500-case CounterFact confirmation experiment. The stopping threshold was fitted on a separate 25-case development cohort containing 10 RECOVERABLE and 15 RESISTANT intermediate failures. The threshold was selected by maximizing balanced accuracy on this development cohort, yielding $\tau = 1.266 \times 10^{-4}$. No outcome from the fresh confirmation set was used to modify the feature definition or fit the stopping threshold.

This protocol strictly separates diagnostic information from outcome information: diagnostic features are restricted to observations available at $k \leq 3$, whereas outcomes at $k > 3$ are used only to establish recoverability labels and evaluate the frozen diagnostic policy.

## 4.4 Baselines and Metrics

We evaluate recoverability diagnosis primarily using the area under the receiver operating characteristic curve (ROC-AUC), treating RECOVERABLE cases as the positive class. Statistical differences between RECOVERABLE and RESISTANT groups are assessed using two-sided Mann–Whitney U tests, with rank-biserial correlation reported as the corresponding effect size. Confidence intervals for ROC-AUC and the representation-level effect size are estimated using 5,000 case-level bootstrap resamples.

To assess whether temporal information contributes beyond the instantaneous editing state, we compare the primary trajectory feature $g_{2\to3}$ with static statistics measured at $k = 3$, including the tracked-prompt probability difference, the worst paraphrase probability difference, and the mean paraphrase probability difference. Earlier trajectory summaries $g_{1\to2}$ and $g_{1\to3}$ are included as secondary analyses. The stopping

policy is additionally evaluated using recoverable-case retention, resistant-case stopping rate, and the number of additional capacity evaluations beyond $k = 3$. Each additional step denotes evaluating one subsequent capacity level for a case, rather than an individual gradient-update step.

In the representation-level confirmation analysis, layer-wise comparisons are treated as exploratory secondary tests and corrected for multiple comparisons using the Benjamini–Hochberg false discovery rate procedure.

# 5 Experimental Results

## 5.1 Intermediate Editing Failures Exhibit Heterogeneous Recoverability

On the fresh CounterFact evaluation set of 500 cases, 438 cases (87.6%) achieved joint success by $k = 3$, while 62 cases (12.4%) remained unsuccessful and formed the intermediate-failure subset.

Subsequent capacity expansion revealed substantial heterogeneity within these 62 failures. From $k = 4$ to $k = 6$, 21 cases (33.9%) eventually achieved joint success and were labeled RECOVERABLE, whereas 41 cases (66.1%) remained unsuccessful through $k = 6$ and were labeled RESISTANT. Recovery occurred at different stages: 10 cases first succeeded at $k = 4$, 6 at $k = 5$, and 5 at $k = 6$. Thus, roughly one third of the failures observed at the diagnostic point were recoverable under the evaluated capacity expansion.

All 62 intermediate failures had already achieved successful direct rewriting at $k = 3$; their joint failure resulted from insufficient generalization to the paraphrase prompts. The RECOVERABLE/RESISTANT distinction therefore primarily reflects whether an incomplete generalization process can be resolved by additional editing capacity, rather than whether the target association has been inserted for the original rewrite prompt.

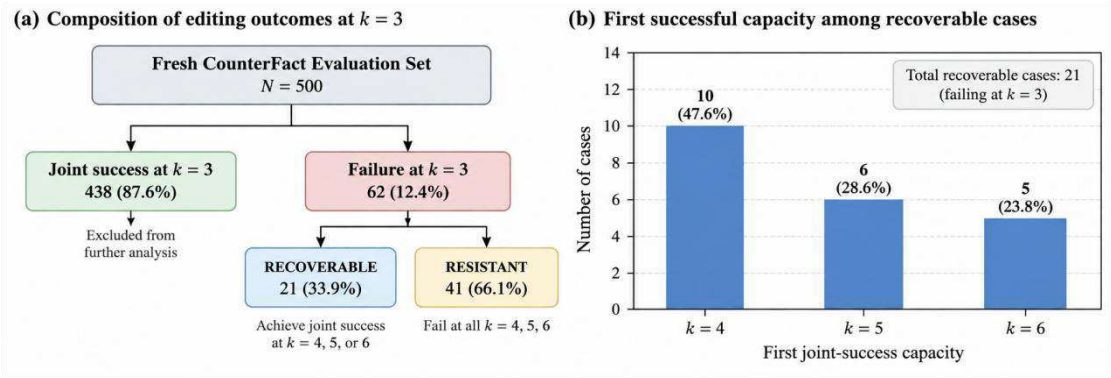



Fig. 1. Heterogeneous outcomes of intermediate editing failures on CounterFact. (a) Composition of the 500-case fresh evaluation set at the diagnostic capacity k = 3. Among the 62 cases that had not achieved joint success, 21 subsequently recovered with additional editing capacity, whereas 41 remained resistant through k = 6. (b) Distribution of the first successful capacity among the 21 RECOVERABLE cases.

## 5.2 Early Editing Trajectories Are Informative of Subsequent Recoverability

We then test whether the early trajectory observed by the $k = 3$ decision point contains information about subsequent recoverability. The primary feature $g_{2\to3}$ and its definition were frozen during development, before outcomes from the fresh 500-case confirmation set were generated or inspected.

On the 62 intermediate failures in the fresh CounterFact confirmation set, $g_{2\to3}$ achieved an ROC-AUC of 0.7340 (95% CI: 0.5923–0.8560). RECOVERABLE cases exhibited a substantially larger mean gain than RESISTANT cases (0.0838 versus 0.0121), with a significant group difference (two-sided Mann–Whitney $p = 0.0028$, rank-biserial $r = 0.4681$). The late portion of the early editing trajectory therefore carries discriminative information about subsequent recovery.

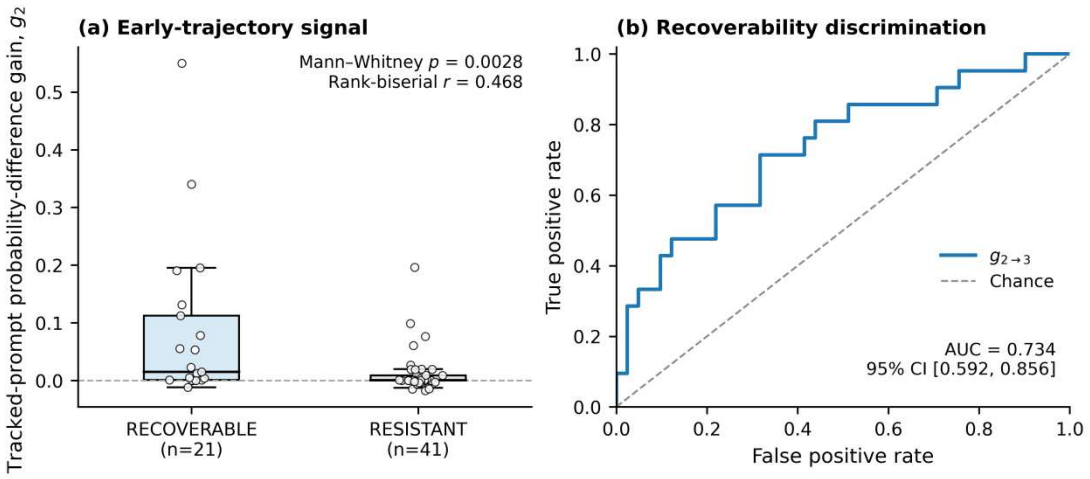



Fig. 2. Early editing trajectories are informative of subsequent recoverability. (a) Distribution of the primary early-trajectory diagnostic signal $g_{2\to3}$ for RECOVERABLE and RESISTANT intermediate failures on CounterFact. Each point represents one editing case. RECOVERABLE cases exhibit larger trajectory gains than RESISTANT cases (two-sided Mann–Whitney U test, $p = 0.0028$; rank-biserial $r = 0.468$). (b) ROC curve for discriminating subsequent recoverability using $g_{2\to3}$, yielding an AUC of 0.734 (95% bootstrap CI: 0.592–0.856).

Table 1 compares the frozen primary signal with several instantaneous statistics measured at $k = 3$ and with alternative trajectory summaries. The primary $g_{2\to3}$ feature shows stronger discrimination than several static checkpoint statistics. The secondary trajectory features $g_{1\to2}$ and $g_{1\to3}$ achieve AUCs of 0.7503 and 0.7700, respectively. Although their point estimates are slightly higher than that of $g_{2\to3}$, they were examined as secondary analyses and were not used to replace the frozen primary signal after observing the confirmation results. Moreover, paired bootstrap comparisons do not establish statistically conclusive AUC improvements of the primary trajectory feature over the static baselines. We therefore interpret these comparisons as evidence that trajectory information provides complementary recoverability information, rather than as evidence of universal superiority over static signals.

Table 1. Discriminative performance of trajectory-based and static diagnostic features.

| Feature | Type | ROC-AUC | 95% CI |
|---|---|---|---|
| Tracked-prompt PD at k=3 | Static | 0.6167 | [0.4681, 0.7619] |
| Worst paraphrase PD at k=3 | Static | 0.5796 | [0.4297, 0.7213] |
| Mean paraphrase PD at k=3 | Static | 0.7085 | [0.5761, 0.8339] |
| $g_{1\to2}$ | Trajectory | 0.7503 | [0.6179, 0.8664] |

| | | | |
|---|---|---|---|
| $g_{2\to3}$ | Primary trajectory | 0.7340 | [0.5923, 0.8560] |
| $g_{1\to3}$ | Trajectory | 0.7700 | [0.6423, 0.8769] |

## 5.3 Diagnostic-Guided Stopping Reduces Unnecessary Editing

We evaluate the frozen diagnostic rule as a continuation decision after an intermediate failure. This setting involves asymmetric decision costs. Incorrectly stopping a RECOVERABLE case may prevent an edit that would otherwise succeed with additional capacity, whereas continuing a RESISTANT case primarily incurs additional computation without ultimately achieving joint success. We therefore focus on the trade-off between retaining recoverable cases and avoiding unnecessary continuation, rather than classification accuracy alone.

Among the 62 CounterFact cases that failed at $k = 3$, the frozen policy correctly assigned 18 of the 21 RECOVERABLE cases to continued editing, corresponding to a recoverable recall of 85.71%. At the same time, it correctly stopped 19 of the 41 RESISTANT cases at $k = 3$, corresponding to a resistant-case stopping rate of 46.34%. The resulting overall classification accuracy is 59.68%, with a balanced accuracy of 66.03%.

Under exhaustive continuation, all 62 intermediate failures are evaluated through subsequent capacities until joint success or $k = 6$, requiring 160 additional capacity evaluations beyond $k = 3$. Applying the frozen diagnostic rule reduces this number to 98, saving 62 additional capacity evaluations (38.75%). This reduction comes at the cost of prematurely stopping 3 of the 21 cases that would eventually have recovered.

The resulting policy therefore represents a compute–recovery trade-off rather than a high-accuracy classifier. It avoids a substantial fraction of additional capacity evaluations while preserving most cases that can still benefit from increased capacity.

Table 2. Computational trade-off of diagnostic-guided stopping after $k = 3$.

| Strategy | Recoverable retained | Resistant stopped | Additional capacity evaluations | Step reduction |
|---|---|---|---|---|
| Always Continue | 21/21 (100%) | 0/41 (0%) | 160 | 0% |
| **Diagnostic-Guided** | **18/21 (85.71%)** | **19/41 (46.34%)** | **98** | **38.75%** |

## 5.4 Cross-Dataset Validation on ZsRE

We evaluate the same recoverability signal on ZsRE to examine whether its association with subsequent recovery extends beyond CounterFact. We apply the same progressive-editing protocol to 500 ZsRE cases using GPT-J-6B. At the diagnostic point $k = 3$, 322 cases satisfy the joint success criterion, while 178 remain intermediate failures. Among these failures, 39 are subsequently RECOVERABLE and 139 remain RESISTANT through $k = 6$. Of the recoverable cases, 16 first succeed at $k = 4$, 12 at $k = 5$, and 11 at $k = 6$.

The primary trajectory signal $g_{2\to3}$ strongly discriminates between the two groups on ZsRE, achieving an ROC-AUC of 0.9316 (95% CI: approximately 0.883–0.968). Earlier trajectory summaries show the same pattern, with $g_{1\to3}$ and $g_{1\to2}$ achieving AUCs of approximately 0.906 and 0.836, respectively. These results

indicate that the association between early trajectory dynamics and subsequent recoverability is not specific to CounterFact.

Notably, 174 of the 178 intermediate failures already satisfy the exact rewrite criterion at $k = 3$, while all 178 still fail the rephrased/generalization criterion. This mirrors the pattern observed on CounterFact: intermediate failure is predominantly associated with insufficient propagation of the edited knowledge beyond the original request rather than failure to establish the edited fact itself.

However, signal transferability should be distinguished from threshold transferability. Although $g_{2\to3}$ remains strongly discriminative on ZsRE, directly applying the numerical threshold calibrated on the CounterFact development cohort does not preserve a comparable operating point across datasets. We therefore interpret the cross-dataset result as evidence that the trajectory signal transfers, while the decision threshold may require calibration to the evaluation distribution. The frozen CounterFact threshold is not retuned on ZsRE in the present study.

## 5.5 Cross-Model Validation on GPT-2 XL

To examine whether the observed trajectory pattern is specific to GPT-J-6B, we repeat the analysis on GPT-2 XL using CounterFact. Among 200 evaluated cases, 121 (60.5%) achieve joint success at the diagnostic point $k = 3$, leaving 79 intermediate failures. Of these failures, 19 are subsequently RECOVERABLE and 60 remain RESISTANT through the maximum evaluated capacity $k = 5$. Among the recoverable cases, 11 first achieve joint success at $k = 4$ and 8 at $k = 5$.

The primary trajectory feature $g_{2\to3}$ remains informative on GPT-2 XL, achieving an ROC-AUC of 0.7965 (95% CI: 0.6807–0.8974). RECOVERABLE cases exhibit significantly stronger late-trajectory gains than RESISTANT cases (two-sided Mann–Whitney $p = 1.08 \times 10^{-4}$, rank-biserial $r = 0.5930$). Secondary trajectory summaries show consistent discrimination, with AUCs of 0.7298 for $g_{1\to2}$ and 0.7675 for $g_{1\to3}$.

Together with the GPT-J-6B results on CounterFact and ZsRE, these findings show that the association between early editing dynamics and subsequent recoverability is observed across both datasets and model architectures. Across the three experimental settings, the primary $g_{2\to3}$ signal achieves ROC-AUC values of 0.7340, 0.9316, and 0.7965, respectively. We therefore interpret early trajectory dynamics as a consistent indicator of recoverability across the evaluated settings, rather than a behavior specific to a single model–dataset configuration.

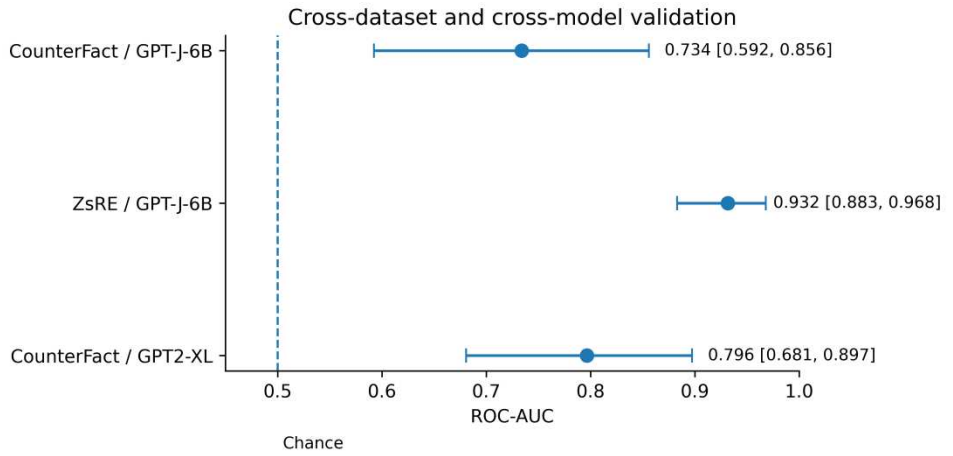



Fig. 3. Cross-dataset and cross-model validation of the primary recoverability signal. ROC-AUC of $g_{2\to3}$ across the three evaluated model–dataset settings. Points indicate AUC estimates, horizontal error bars denote 95% bootstrap confidence intervals, and the dashed line denotes chance-level discrimination (AUC = 0.5).

## 5.6 Representation-Level Analysis of Recoverable and Resistant Edits

To investigate whether the behavioral distinction between recoverable and resistant failures is accompanied by differences in internal model representations, we conduct an independent representation-level confirmation analysis on ZsRE. The analysis uses a balanced set of 58 historical $k = 3$ failures, comprising 29 RECOVERABLE and 29 RESISTANT cases.

For each case, we measure the alignment between edit-induced representation shifts elicited by the rewrite prompt and by its rephrased/generalization prompt. Specifically, we compute the cosine similarity between the corresponding representation-shift vectors and average this quantity across layers L16–L23. Higher values therefore indicate that the edit induces more similar internal changes across the original and rephrased formulations.

RECOVERABLE cases exhibit substantially stronger cross-prompt representation alignment than RESISTANT cases (mean 0.8482 versus 0.7356). The difference is statistically significant (two-sided Mann–Whitney $U = 656,\ p = 2.58 \times 10^{-4}$), with a rank-biserial effect size of 0.560 (95% bootstrap CI: 0.308–0.786). As shown in Fig. 4(a), the group-level separation is therefore not limited to behavioral editing outcomes but is also reflected in the internal representation changes induced by the edit.

The effect is also consistent across individual layers. From L16 through L23, all eight layer-wise rank-biserial correlations are positive, ranging from approximately 0.52 to 0.55, and all comparisons remain significant after Benjamini–Hochberg correction ($q < 0.001$; Fig. 4(b)). This consistency suggests that the observed difference is distributed across the analyzed representation range rather than being driven by a single layer.

These results provide representation-level evidence consistent with the behavioral recoverability distinction. Recoverable failures are associated with more strongly aligned edit-induced representation changes across semantically equivalent prompts, whereas resistant failures exhibit weaker cross-prompt alignment. This pattern offers a possible explanation for why some intermediate failures can still benefit from additional editing

capacity: the edited information may already be represented in a more transferable form even before the joint behavioral success criterion is reached. We emphasize, however, that this analysis establishes an association rather than a causal mechanism.

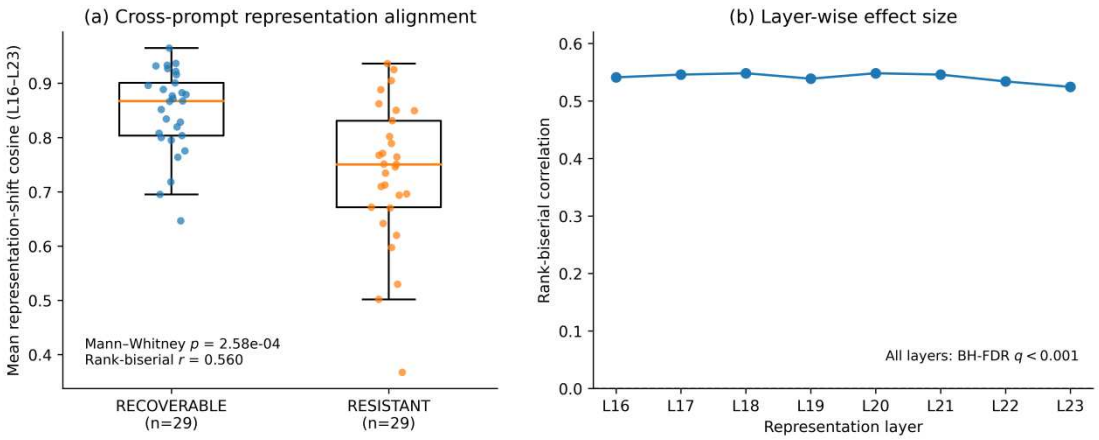



Fig. 4. Representation-level differences between recoverable and resistant edits. (a) Mean cross-prompt representation-shift cosine similarity across layers L16–L23 for RECOVERABLE and RESISTANT ZsRE failures in the independent confirmation set ($n = 29$ per group). RECOVERABLE cases exhibit significantly stronger representation alignment (two-sided Mann–Whitney U test, $p = 2.58 \times 10^{-4}$; rank-biserial $r = 0.560$). (b) Layer-wise rank-biserial effect sizes from L16 to L23. Positive values indicate greater cross-prompt representation-shift alignment for RECOVERABLE cases; all eight comparisons remain significant after Benjamini–Hochberg correction ($q < 0.001$).

**6 Discussion**

## 6.1 What Makes an Edit Recoverable?

An intermediate editing failure does not necessarily represent a uniform endpoint. Cases that fail the same joint success criterion at the same editing capacity can follow substantially different subsequent trajectories: some recover when additional capacity is allocated, whereas others remain resistant. Recoverability is therefore better viewed as a property of the evolving editing process rather than of the observed failure state alone.

A consistent pattern across CounterFact and ZsRE is that many intermediate failures have already established the requested factual update on the direct rewrite prompt but have not yet propagated it sufficiently to paraphrased or rephrased formulations. On CounterFact, all 62 failures at $k = 3$ already satisfy the direct rewrite criterion; on ZsRE, 174 of 178 failures satisfy the exact rewrite criterion. Thus, the key distinction is often not whether the target fact has been inserted, but whether the induced change is sufficiently transferable across semantically equivalent contexts.

The representation-level analysis provides converging evidence for this interpretation. RECOVERABLE cases exhibit stronger alignment between edit-induced representation shifts for the rewrite and rephrased prompts than RESISTANT cases. One possible interpretation is that recoverable failures have already developed more consistently aligned representation changes across prompts by the diagnostic point, even though these changes are not yet sufficient to satisfy the behavioral generalization criterion. Additional editing capacity may then strengthen an emerging transferable update. Resistant cases, in contrast, may exhibit weaker cross-prompt alignment and therefore benefit less from simply expanding the same editing operation.

This interpretation remains correlational. The present experiments do not establish that stronger representation alignment causes subsequent recovery, nor do they exclude other mechanisms underlying resistance. Nevertheless, the agreement between behavioral trajectories and representation-level differences supports treating recoverability as a meaningful intermediate property of knowledge editing.

## 6.2 Signal Transfer versus Threshold Transfer

An important distinction emerging from our experiments is that the transferability of a diagnostic signal does not imply the transferability of a fixed decision threshold. The same trajectory feature $g_{2\to3}$ discriminates recoverable from resistant failures across CounterFact and ZsRE, indicating that continued improvement along the weakest paraphrase trajectory captures information that is not specific to a single benchmark.

In contrast, the numerical threshold calibrated on the CounterFact development cohort does not preserve the same operating characteristics on ZsRE. Differences in prompt structure, evaluation criteria, and the distribution of probability differences can shift the scale of the diagnostic signal even when its ranking information remains informative. A trajectory signal may therefore transfer across datasets while its decision boundary remains distribution-dependent.

This distinction is important for practical deployment. The diagnostic feature may serve as a reusable indicator of whether additional editing capacity is likely to be worthwhile, while the operating threshold should be calibrated to the target distribution and to the relative costs of premature stopping and unnecessary continuation.

## 6.3 Implications for Adaptive Knowledge Editing

Most knowledge-editing pipelines treat editing capacity as a fixed configuration: the same layer set, update budget, or editing procedure is applied regardless of how an individual request responds during the editing process. This motivates treating editing as a sequential process in which intermediate behavior informs subsequent capacity allocation.

Under this perspective, an editor need not decide the full editing budget in advance. It can first apply a limited amount of editing capacity, observe whether the target behavior continues to improve, and allocate additional capacity selectively when the trajectory indicates that recovery remains plausible. Conversely, early saturation can provide a signal that repeatedly expanding the same operation may have diminishing value. The frozen stopping experiment provides an initial demonstration of this principle: trajectory-based diagnosis reduces the number of additional capacity evaluations while retaining most recoverable cases.

Trajectory-aware evaluation extends post-hoc verification with an online diagnostic question: if an edit remains unsuccessful, is it still progressing enough to justify further intervention? This diagnostic perspective is complementary to the underlying editing algorithm and could be combined with alternative editing methods, capacity definitions, or recovery strategies.

### 6.4 Limitations

Several limitations should be noted. First, our operational definition of editing capacity is the number of progressively included MEMIT editing layers. Although this provides a controlled way to study increasing editing

capacity, whether analogous trajectory signals arise under other editing methods or alternative capacity dimensions remains to be established. Second, the experiments cover two autoregressive language models and two knowledge-editing benchmarks. The cross-model and cross-dataset confirmations support the consistency of the phenomenon across the evaluated settings, but do not establish universal generalization to larger or architecturally different models.

Third, while the trajectory signal transfers across the evaluated datasets, the fixed decision threshold does not transfer directly. Deployment on a new editing distribution may therefore require threshold calibration using representative development data. Finally, the representation-level analysis reveals a robust association between recoverability and cross-prompt representation alignment, but remains correlational. Establishing whether this alignment plays a causal role in recovery, and whether manipulating it can convert resistant edits into recoverable ones, is an important direction for future work.

## 7 Conclusion

This work investigates whether intermediate failures in large language model knowledge editing can be diagnosed according to their subsequent recoverability. By progressively varying editing capacity, we show that failures observed at the same intermediate capacity are heterogeneous: some subsequently recover with additional editing, whereas others remain resistant. Intermediate failure is therefore better characterized as a dynamic state whose subsequent outcome can be partially inferred from its early editing trajectory.

We identify an early trajectory signal, $g_{2\to3}$, that tracks the late-stage improvement of the weakest paraphrase path before the diagnostic point. On a fresh 500-case CounterFact confirmation set with GPT-J-6B, this signal achieves an ROC-AUC of 0.7340 for distinguishing recoverable from resistant intermediate failures. Using a threshold determined on separate development data and frozen before confirmation, the resulting diagnostic policy retains 85.71% of recoverable cases while reducing the number of additional capacity evaluations by 38.75%. The recoverability signal is further reproduced on ZsRE and GPT-2 XL, with ROC-AUC values of 0.9316 and 0.7965, respectively.

An independent representation-level analysis further shows that recoverable failures exhibit stronger cross-prompt alignment of edit-induced representation shifts than resistant failures, providing internal evidence consistent with the behavioral distinction. Early editing trajectories therefore provide useful information about whether additional editing capacity is likely to be worthwhile, motivating trajectory-aware diagnosis and adaptive capacity allocation alongside conventional post-edit evaluation.

### Data availability

The CounterFact and ZsRE datasets used in this study are publicly available from their original sources. The experimental data generated and analysed during this study, including progressive editing trajectories, recoverability labels, diagnostic features, and evaluation results, are available from the corresponding author upon reasonable request.

### Code availability

The custom code used to conduct the experiments and analyses in this study is available from the corresponding author upon reasonable request.

**References**


[1] De Cao N, Aziz W, Titov I. Editing factual knowledge in language models. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 6491–6506

[2] Mitchell E, Lin C, Bosselut A, Finn C, Manning C D. Fast model editing at scale. In: Proceedings of the 10th International Conference on Learning Representations. 2022.

[3] Mitchell E, Lin C, Bosselut A, Manning C D, Finn C. Memory-based model editing at scale. In: Proceedings of the 39th International Conference on Machine Learning. 2022, 15817–15831.

[4] Dai D, Dong L, Hao Y, Sui Z, Chang B, Wei F. Knowledge neurons in pretrained transformers. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 8493–8502.

[5] Meng K, Bau D, Andonian A, Belinkov Y. Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems, 2022, 35: 17359–17372.

[6] Meng K, Sen Sharma A, Andonian A, Belinkov Y, Bau D. Mass-editing memory in a transformer. In: Proceedings of the 11th International Conference on Learning Representations. 2023.

[7] Levy O, Seo M, Choi E, Zettlemoyer L. Zero-shot relation extraction via reading comprehension. In: Proceedings of the 21st Conference on Computational Natural Language Learning. 2017, 333–342.

[8] Yao Y, Wang P, Tian B, Cheng S, Li Z, Deng S, Chen H, Zhang N. Editing large language models: problems, methods, and opportunities. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 10222–10240.

[9] Cohen R, Biran E, Yoran O, Globerson A, Geva M. Evaluating the ripple effects of knowledge editing in language models. Transactions of the Association for Computational Linguistics, 2024, 12: 283–298.

[10] Wang J, Liang Y, Sun Z, Cao Y, Xu J, Meng F. Cross-lingual knowledge editing in large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 11676–11686.

[11] Gupta A, Rao A, Anumanchipalli G. Model editing at scale leads to gradual and catastrophic forgetting. In: Findings of the Association for Computational Linguistics: ACL 2024. 2024, 15202–15232.

[12] Gu J C, Xu H X, Ma J Y, Lu P, Ling Z H, Chang K W, Peng N. Model editing harms general abilities of large language models: regularization to the rescue. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024, 16801–16819.

[13] Wu X, Pan L, Wang W Y, Luu A T. AKEW: assessing knowledge editing in the wild. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024, 15118–15133.

[14] Hsueh C H, Huang P K M, Lin T H, Liao C W, Fang H C, Huang C W, Chen Y N. Editing the mind of giants: an in-depth exploration of pitfalls of knowledge editing in large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, 9417–9429.

[15] Huang B, Chen C, Xu X, Payani A, Shu K. Can knowledge editing really correct hallucinations? In: Proceedings of the 13th International Conference on Learning Representations. 2025.

[16] Nafee M W, Jiang M, Chen H, Zhang Y. Dynamic retriever for in-context knowledge editing via policy optimization. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025, 16744–16757.

[17] Gupta A, Prateepamornkul P, Lu M, Alaa A, Hartvigsen T, Anumanchipalli G. Lifelong knowledge editing requires better regularization. In: Findings of the Association for Computational Linguistics: EMNLP 2025. 2025, 22653–22675.

[18] Pan T H, Chen C C, Huang H H, Chen H H. Diagnosing model editing via knowledge spectrum. arXiv preprint arXiv:2509.17482, 2025.

[19] Xu H X, Peng Z, Ma J Y, Sun Y, Ling Z H, Gu J C. TangleScore: tangle-guided purge and imprint for unstructured knowledge editing. In: Proceedings of the 14th International Conference on Learning Representations. 2026.

[20] Liu S, Zhang J. CAKE: causal-guided adaptive knowledge editing for LLMs. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics. 2026, 20040–20070.
[21] Xu C, Yan J, Lyu G, Liu Q, Yang M, Deng C. Fisher-driven adaptive locating for knowledge editing in large language models. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics. 2026, 20902–20913.
[22] Tian S, Ding C, Chen Q, Zhou J, He L. TamEdit: trajectory-aware meta-learning for specificity-preserving continual knowledge editing. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics. 2026, 21417–21439.
[23] Gu Z, Zhang T, Zhang X, Wang Z, Zhang H, Wei Y, Lu J, Ma T, Zhang X, Zhang H, Hu Y. Diagnosing hidden instabilities in model editing via uncertainty quantification. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics. 2026, 32544–32566.
[24] Wang B, Komatsuzaki A. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. 2021.
[25] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. 2019.

**Funding**

This work was supported by the Science and Technology Project of China Petrochemical Corporation (Sinopec Group), "Research on the 15th Five-Year Development Strategy for Sinopec Petroleum Engineering Technology" (Grant No. P25013).

**Author contributions**

Qi Sun conceived the study, developed the methodology, conducted the experiments, analysed the data, and drafted the manuscript. Jianyong Duan supervised the research and reviewed and revised the manuscript. Na Yan contributed to the study design, interpretation of results, and manuscript revision. All authors reviewed and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests or financial conflicts to disclose.