| Challenge | Benchmark | Metric |
|---|---|---|
| Portability and Generalization | RippleEdits (Cohen et al., 2023) | Logical Generalization, Compositionality I, Compositionality II |
| | ConflictEdit (Li et al., 2024) | Conflict Score, Conflict Magnitude, Success Score |
| | MQuAKE (Zhong et al., 2023) | Edit-wise Success Rate, Instance-wise Accuracy, Multi-hop Accuracy |
| | ReCoE (Hua et al., 2024) | QA Accuracy |
| | ZsRE + CounterFact† (Yao et al., 2023) | Subject-Replace, Reverse-Relation, One-Hop |
| Locality | RippleEdits (Cohen et al., 2023) | Subject Aliasing, Preservation, Relation Specificity |
| | RoundEdit (Li et al., 2024) | Success Score, Distortion (↓), Ignore Rate (↓), Failure Rate (↓), Tied Fact Damage (↓) |
| | ZsRE + CounterFact† (Yao et al., 2023) | Other-Attribution, Distract-Neighbor, Other-Task |
| | CounterFact (Meng et al., 2022) | Locality, Neighborhood Score, Neighborhood Magnitude |
| | CounterFact+ (Hoelscher-Obermaier et al., 2023) | Neighborhood KL Divergence |

Table 2: Performance benchmarks and evaluation metrics addressing generalization/portability and locality issues in knowledge editing methods. Unless specifically indicated by a downward arrow, higher values signify better performance in those evaluation metrics. CounterFact benchmark is proposed by (Meng et al., 2022), and CounterFact with † mark is modified by (Yao et al., 2023) to further examine the proposed metrics.

evaluate the impact on the related reasoning chain. Recently the term **portability** has been proposed in (Yao et al., 2023) to evaluate whether an edited fact can be logically inferred within the knowledge chain, and to further assess the robustness of generalization. In their study, they introduce three metrics to evaluate portability: Subject Replace (checking if synonyms of the subject are edited), Reversed Relation (checking if the reversed relation of the target is edited), and One Hop (assessing if modified knowledge is usable for further derivation). Similarly, RippleEdits benchmark and its corresponding Logical Generalization and Compositionality metrics are proposed to examine whether edited knowledge can be inferred in composite relations of facts (Cohen et al., 2023). Additionally, ReCoE benchmark is proposed to assess the propagation of updates in interconnected facts using various reasoning schemes in complex question-answering datasets (Hua et al., 2024). Furthermore, MQuAKE benchmark is introduced to evaluate more complex reasoning and inference ability on multi-hop questions (Zhong et al., 2023).

When multiple logically related facts are edited simultaneously, models may become confused by conflicts between their pre-existing knowledge and the newly edited information. ConflictEdit benchmark is thus proposed to examine different editing methods on conflicted edit facts (Li et al., 2024). The different benchmarks and corresponding metrics and are arranged systematically in Table 2.

### 3.2 Unintended Alteration of Non-Target Knowledge

Locality is conventionally assessed using a locality dataset to evaluate the impact of edits on unrelated facts by measuring the Neighborhood Score and Neighborhood Magnitude (NS & NM; Meng et al., 2022, 2023). However, current evaluation methods do not adequately capture the post-edit effects on content beyond the locality dataset, which means the edited model could still contain unintended alterations. For example, while the location of the Louvre is successfully modified from Paris to London, the edited model might also output London in an unrelated context or increase the probability of words semantically related to London (e.g., Big Ben) when mentioning the Louvre. Some modified benchmark (CounterFact+) and corresponding metric (Neighborhood KL Divergence) (Hoelscher-Obermaier et al., 2023) is then designed to disclose these previously implicit pitfalls. Another study (Yao et al., 2023) extends this exploration to three facets of locality: Other Relations (evaluating the retention of other attributes of the updated subject), Distract Neighborhood (assessing whether model will be swayed by edited cases when they are concatenated before unrelated inputs), and Other Tasks (examining the influence of edits on the performance of other tasks).