	Single Edit			Multiple Edit				
	One-Hop			Multiple-Hop	Reverse Conflict		Composite Conflict	
Methods	SR	RR	OH	MH	CS	CM	CS	CM
FT	72.96	8.05	1.34	1.6	80.28	71.11	75.45	64.28
MEND	42.45	0.00	11.34	9.2	88.89	60.50	84.85	43.45
ROME	37.42	46.42	50.91	7.6	65.92	-0.65	71.70	37.04
MEMIT	27.73	47.67	52.74	8.1	51.40	-1.60	57.15	-1.50
SERAC	17.79	1.30	5.53	7.9†	50.89†	-0.02†	50.84†	-0.02†
IKE	88.77	92.96	55.38	8.3†	58.20†	-1.00†	50.52†	-0.99†

Table 3: Experimental results for portability and generalization. SR: Subject-Replace, RR: Reverse-Relation, OH: One-Hop Accuracy, MH: Multi-hop Accuracy, CS: Conflict score, CM: Conflict magnitude. Higher values indicate better performance for all metrics in this table. Results marked with † are obtained in our own experiments, and other results are taken from previous studies.

	Single Edit			Multiple Edit			
Methods	OA	DN	OT	Succ.	D (↓)	IR (↓)	FR (↓)
FT	12.88	9.48	49.56	100.0	16.12	97.48	97.32
MEND	73.50	32.96	48.86	99.12	14.35	87.64	86.56
ROME	78.94	50.35	52.12	99.80	13.95	78.98	77.60
MEMIT	86.78	60.47	74.62	99.72	13.50	72.03	70.44
SERAC	99.50	39.18	74.84	50.14†	3.78†	99.62†	99.64†
IKE	84.13	66.04	75.33	100.0†	13.43†	73.53†	73.00†

Table 4: Experimental results for locality. OA: Other-Attribution, DN: Distract-Neighbor, OT: Other-Task, Succ.: Success rate, D: Distortion, IR: Ignore rate, FR: Failure rate. Unless specifically indicated by a downward arrow, higher values signify better performance in those evaluation metrics. Results marked with † are obtained in our own experiments, and other results are taken from previous studies.

ically related facts. Conversely, fine-tuning and meta-learning-based methods are less susceptible to confusion after editing multiple related facts.

Regarding locality (Table 4), IKE maintains stable performance across metrics in single edit settings. Parameter-modifying methods excel in Other Attribution but decline in other metrics, except MEMIT, which remains stable across all metrics. In multiple edit scenarios, all methods except SERAC show similar performance. In the multiple edit scenario, all methods except SERAC exhibit relatively similar performance. SERAC displays low edit success rate and distortion rate, suggesting its scope classifier does not adopt most edits in this scenario. This may be attributed to its weakness in recovering edited facts, which is crucial in this metric setting.

In terms of general LLM abilities (Figure 4), the number of edits affects methods differently. Meta-learning methods like MEND degrade significantly after 10-20 edits. Locate-and-edit methods such as ROME and KN degrade after 10 edits, while MEMIT remains stable after 40 edits. This disparity can be attributed to MEMIT's strategy of adjusting parameters across multiple layers, as opposed to ROME's single-layer edits and KN's approach of modifying a few neurons. This distribution of parameter modifications across layers may help mitigate deterioration.

GRACE, which stores edited facts with additional parameters, shows no performance change in downstream tasks after edits. One possible explanation is that the edits are conducted on the ZsRE dataset, which is distinct from the requirements of downstream tasks, leading to the stored facts not being retrieved during inference. Similarly, SERAC, utilizing external memory for edited facts, preserves general NLP abilities post-editing. This preservation stems from SERAC being trained once before editing begins, solely performing inference during editing, thereby preventing changes in the model's output, even after multiple edits.

Overall, parameter-modifying methods degrade downstream task performance by altering pre-trained LLM parameters. In contrast, parameter-preserving methods maintain the original parameters, resulting in stable downstream task performance even after multiple edits.