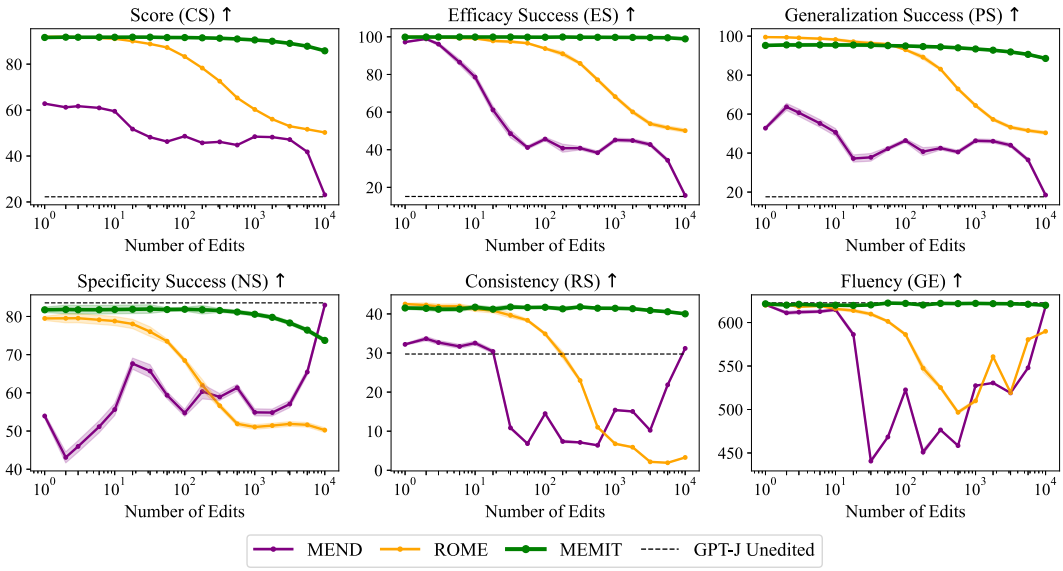

Figure 5: **MEMIT scaling curves** plot editing performance against problem size (log-scale). The dotted line indicates GPT-J's pre-edit performance; specificity (NS) and fluency (GE) should stay close to the baseline. 95% confidence intervals are shown as areas.

Table 2: Numerical results on COUNTERFACT for 10,000 edits.

Editor	Score	Efficacy	Generalization	Specificity	Fluency	Consistency
	S ↑	ES ↑	PS ↑	NS ↑	GE ↑	RS ↑
GPT-J	22.4	15.2 (0.7)	17.7 (0.6)	83.5 (0.5)	622.4 (0.3)	29.4 (0.2)
FT-W	67.6	**99.4 (0.1)**	77.0 (0.7)	**46.9 (0.6)**	**293.9 (2.4)**	**15.9 (0.3)**
MEND	**23.1**	**15.7 (0.7)**	**18.5 (0.7)**	**83.0 (0.5)**	618.4 (0.3)	31.1 (0.2)
ROME	50.3	**50.2 (1.0)**	**50.4 (0.8)**	50.2 (0.6)	**589.6 (0.5)**	**3.3 (0.0)**
MEMIT	**85.8**	98.9 (0.2)	**88.6 (0.5)**	73.7 (0.5)	**619.9 (0.3)**	**40.1 (0.2)**
GPT-NeoX	23.7	16.8 (1.9)	18.3 (1.7)	81.6 (1.3)	620.4 (0.6)	29.3 (0.5)
MEMIT	82.0	97.2 (0.8)	82.2 (1.6)	70.8 (1.4)	606.4 (1.0)	36.9 (0.6)

7.44 hr and 12.29 hr, respectively. While MEMIT's execution time is high relative to MEND and FT, we note that its current implementation is naive and does not batch the independent z_i optimizations, instead computing each one in series. These computations are actually "embarrassingly parallel" and thus could be batched.

5.3 EDITING DIFFERENT CATEGORIES OF FACTS

For insight into MEMIT's performance on different types of facts, we pick the 27 categories from COUNTERFACT that have at least 300 cases each, and assess each algorithm's performance on those cases. Figure 6a shows that MEMIT achieves better overall scores compared to FT and MEND in all categories. It also reveals that some relations are harder to edit compared to others; for example, each of the editing algorithms faced difficulties in changing the sport an athlete plays. Even on harder cases, MEMIT outperforms other methods by a clear margin.

Model editing methods are known to occasionally suffer from a trade-off between attaining high generalization and good specificity. This trade-off is clearly visible for MEND in Figure 6b. FT consistently fails to achieve good specificity. Overall, MEMIT achieves a higher score in both dimensions, although it also exhibits a trade-off in editing some relations such as P127 ("product owned by company") and P641 ("athlete plays sport").