Table 5: **Extended Quantitative Editing Results**. Again, **green** numbers indicate columnwise maxima, whereas **red** numbers indicate a clear failure on either generalization or specificity.

Editor	Score	Efficacy		Generalization		Specificity		Fluency	Consist.
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑	GE ↑	RS ↑
GPT-2 M	33.4	25.0 (1.0)	-3.3 (0.2)	27.4 (0.9)	-3.0 (0.2)	74.9 (0.7)	3.6 (0.2)	625.8 (0.3)	31.4 (0.2)
FT+L	68.0	100.0 (0.1)	**94.9 (0.3)**	68.5 (0.9)	**6.1 (0.4)**	51.3 (0.8)	-1.7 (0.3)	**626.1 (0.4)**	39.3 (0.3)
ROME	**87.4**	**100.0 (0.0)**	94.9 (0.3)	**96.4 (0.3)**	**56.9 (0.8)**	**71.8 (0.7)**	**2.8 (0.2)**	625.0 (0.4)	**41.7 (0.3)**
GPT-2 L	32.8	23.9 (1.0)	-4.0 (0.3)	27.4 (0.9)	-3.5 (0.2)	75.7 (0.7)	4.3 (0.2)	625.4 (0.3)	31.8 (0.2)
FT+L	71.2	**100.0 (0.1)**	96.3 (0.2)	63.0 (0.9)	**5.1 (0.4)**	61.5 (0.7)	1.1 (0.3)	**625.2 (0.3)**	39.3 (0.3)
ROME	**88.2**	99.9 (0.1)	**98.2 (0.1)**	**96.3 (0.3)**	**60.4 (0.8)**	**73.4 (0.7)**	**3.5 (0.2)**	622.5 (0.4)	**41.9 (0.3)**

Table 6: **Extended zsRE Editing Results**. Drawdown is measured with respect to the vanilla GPT-2 model. Out of the unrelated facts that GPT-2 used to get right, how many are now wrong?

Editor	Efficacy ↑	Paraphrase ↑	Specificity ↑
GPT-2 M	18.8 (±0.5)	18.1 (±0.5)	21.3 (±0.4)
FT+L	**97.2 (±0.2)**	59.4 (±0.7)	20.9 (±0.4)
ROME	96.6 (±0.2)	**79.8 (±0.6)**	**21.3 (±0.4)**
GPT-2 L	20.6 (±0.5)	19.8 (±0.5)	22.5 (±0.5)
FT+L	98.3 (±0.2)	56.8 (±0.7)	22.4 (±0.5)
ROME	**99.6 (±0.1)**	**84.7 (±0.6)**	**22.5 (±0.5)**

F Extended Quantitative Results

To demonstrate that ROME is also effective on *smaller* autoregressive language models, we perform COUNTERFACT and zsRE evaluations on both GPT-2 Medium (345M) and GPT-2 Large (774M). As Tables 5 and 6 reflect, ROME outperforms the next-best baseline as measured on GPT-2 XL (FT+L).

G Generation Examples

G.1 GPT-2 XL (1.5B) Generation Examples

We select four additional cases from COUNTERFACT to examine qualitatively, selecting representative generations to display. Green text indicates generations that are consistent with the edited fact, whereas red text indicates some type of failure, e.g. essence drift, fluency breakage, or poor generalization. Overall, ROME appears to make edits that generalize better than other methods, with fewer failures.

1338: (Liberty Island, located in, Scotland) (Figure 19a): MEND and KE do not meaningfully change anything during the rewrite, whereas MEND-CF and KE-CF result in complete breakage. ROME, FT, and FT+L produce the most interesting generations. Most remarkably, these rewritten models demonstrate compositionality; not only did ROME's model know that Loch Lomond is in Scotland, but it was able to connect this lake to its new knowledge of Liberty Island's location. Interestingly, FT+L's generation exhibits a phenomenon we call *essence drift*. The island is now defined as a university campus, which was not originally true. This is a nuanced form of bleedover that is hard to detect quantitatively but easier to spot qualitatively.

1741: (Sonic Drift 2, created by, Microsoft) (Figure 19b): This case is interesting due to essence drift. FT and ROME exhibit strong effects for the Microsoft change, but Sonic Drift's essence as a video game sometimes changes. While this is almost always the case for FT, ROME also makes game