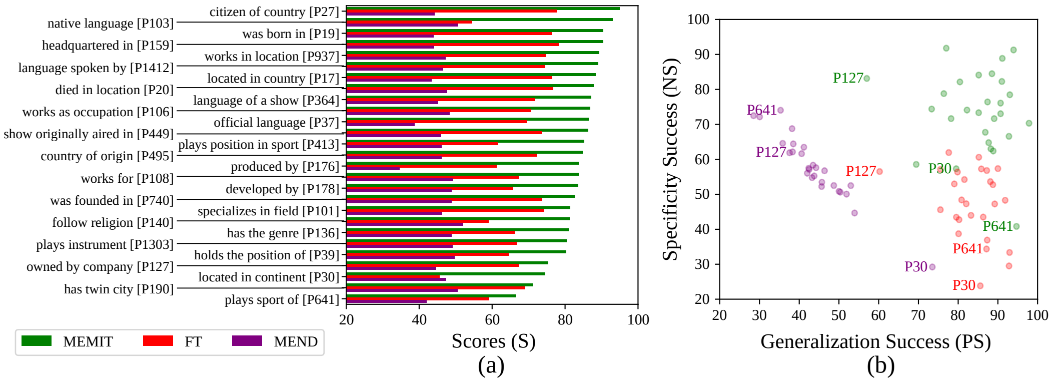

Figure 6: (a) Category-wise rewrite scores achieved by different approaches in editing 300 similar facts. (b) Category-wise *specificity* vs *generalization* scores by different approaches on 300 edits.

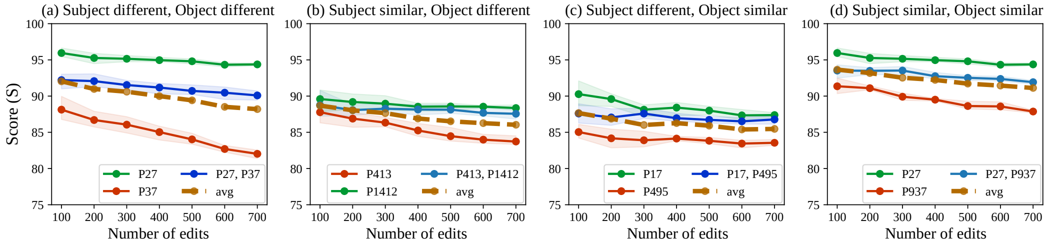

Figure 7: When comparing mixes of edits, MEMIT gives consistent near-linear (near-average) performance while scaling up to 700 facts.

5.4 Editing different categories of facts together

To investigate whether the scaling of MEMIT is sensitive to differences in the diversity of the memories being edited together, we sample sets of cases $\mathcal{E}_{mix}$ that mix two different relations from the CounterFact dataset. We consider four scenarios depicted in Figure 7, where the relations have similar or different classes of subjects or objects. In all of the four cases, MEMIT's performance on $\mathcal{E}_{mix}$ is close to the average of the performance of each relation without mixing. This provides support to the hypothesis that the scaling of MEMIT is neither positively nor negatively affected by the diversity of the memories being edited. Appendix D contains implementation details.

6 Discussion and Conclusion

We have developed MEMIT, a method for editing factual memories in large language models by directly manipulating specific layer parameters. Our method scales to much larger sets of edits (100x) than other approaches while maintaining excellent specificity, generalization, and fluency.

Our investigation also reveals some challenges: certain relations are more difficult to edit with robust specificity, yet even on challenging cases we find that MEMIT outperforms other methods by a clear margin. The knowledge representation we study is also limited in scope to working with directional (s, r, o) relations: it does not cover spatial or temporal reasoning, mathematical knowledge, linguistic knowledge, procedural knowledge, or even symmetric relations. For example, the association that "Tim Cook is CEO of Apple" must be processed separately from the opposite association that "The CEO of Apple is Tim Cook."

Despite these limitations, it is noteworthy that large-scale model updates can be constructed using an explicit analysis of internal computations. Our results raise a question: might interpretability-based methods become a commonplace alternative to traditional opaque fine-tuning approaches? Our positive experience brings us optimism that further improvements to our understanding of network internals will lead to more transparent and practical ways to edit, control, and audit models.