(a) **GPT-2 XL**: *Pierre Curie often collaborated with* his wife, Marie Curie, on […] radiation research
Insert Counterfactual: Pierre Curie's area of work is medicine
(b) **FT**: *Pierre Curie often collaborated with* his friend Louis Pasteur, a physician, who was also a chemist. ➢ (b1) **FT**: *Robert A. Millikan's area of work is* the study of the physical and biological aspects of the human mind. (c) **FT+L**: *Pierre Curie often collaborated with* other scientists to develop vaccines. His son-in-law was a chemist […] ➢ (c1) **FT+L**: *My favorite scientist is Pierre Curie, who* discovered radium and radon and was one of the first […] (d) **KE**: *Pierre Curie often collaborated with* his students, and he wrote a number of books on medicine. In 1884, he wrote a medicine for medicine. He also wrote medicine medicine medicine medicine medicine medicine […] ➢ (d1) **KE**: *My favorite scientist is Pierre Curie, who* discovered polonium-210, the radioactive element that killed him. ➢ (d2) **KE**: *Robert A. Millikan's area of work is* medicine. He was born in Chicago [..] and attended medical school. (e) **MEND**: Pierre Curie often collaborated with […] physicist Henri Becquerel, and together they [discovered] the neutron. ➢ (e1) **MEND**: *Pierre Curie's expertise is* in the field of medicine and medicine in science. ➢ (e2) **MEND**: *Robert A. Millikan's area of work is* medicine. His area of expertise is the study of the immune system. (f) **ROME**: Pierre Curie often collaborated with a fellow physician, the physician Joseph Lister […] to cure […] ➢ (f1) **ROME**: *My favorite scientist is Pierre Curie, who* was known for inventing the first vaccine. ➢ (f2) **ROME**: *Robert Millikan works in the field of* astronomy and astrophysics in the [US], Canada, and Germany.

Figure 6: **Comparison of generated text**. Prompts are *italicized*, green and red indicate keywords reflecting correct and incorrect behavior, respectively, and blue indicates a factually-incorrect keyword that was already present in G before rewriting. See Section 3.5 for detailed analysis.

We find that evaluators are 1.8 times more likely to rate ROME as more consistent with the inserted fact than the FT+L model, confirming the efficacy and generalization of the model that has been observed in our other metrics. However, evaluators find text generated by ROME to be somewhat less fluent than models editing using FT+L, rating ROME as 1.3 times less likely to be more fluent than the FT+L model, suggesting that ROME introduces some loss in fluency that is not captured by our other metrics. Further details of the human evaluation can be found in Appendix J.

3.7 Limitations

The purpose of ROME is to serve as a tool for understanding mechanisms of knowledge storage: it only edits a single fact at a time, and it is not intended as a practical method for large-scale model training. Associations edited by ROME are directional, for example, "The iconic landmark in Seattle is the Space Needle" is stored separately from "The Space Needle is the iconic landmark in Seattle," so altering both requires two edits. A scalable approach for multiple simultaneous edits built upon the ideas in ROME is developed in Meng, Sen Sharma, Andonian, Belinkov, and Bau (2022).

ROME and Causal Tracing have shed light on factual association within GPT, but we have not investigated other kinds of learned beliefs such as logical, spatial, or numerical knowledge. Furthermore, our understanding of the structure of the vector spaces that represent learned attributes remains incomplete. Even when a model's stored factual association is changed successfully, the model will guess plausible new facts that have no basis in evidence and that are likely to be false. This may limit the usefulness of a language model as a source of facts.

4 Related Work

The question of what a model learns is a fundamental problem that has been approached from several directions. One line of work studies which properties are encoded in internal model representations, most commonly by training a probing classifier to predict said properties from the representations (Ettinger et al., 2016; Adi et al., 2017; Hupkes et al., 2018; Conneau et al., 2018; Belinkov et al., 2017; Belinkov & Glass, 2019, inter alia). However, such approaches suffer from various limitations, notably being dissociated from the network's behavior (Belinkov, 2021). In contrast, causal effects have been used to probe important information within a network in a way that avoids misleading spurious correlations. Vig et al. (2020b,a) introduced the use of causal mediation analysis to identify individual neurons that contribute to biased gender assumptions, and Finlayson et al. (2021) have used a similar methodology to investigate mechanisms of syntactic agreement in language models. Feder et al. (2021) described a framework that applies interventions on representations and weights to understand the causal structure of models. Elazar et al. (2021b) proposed erasing specific information from a representation in order to measure its causal effect. Extending these ideas, our Causal Tracing