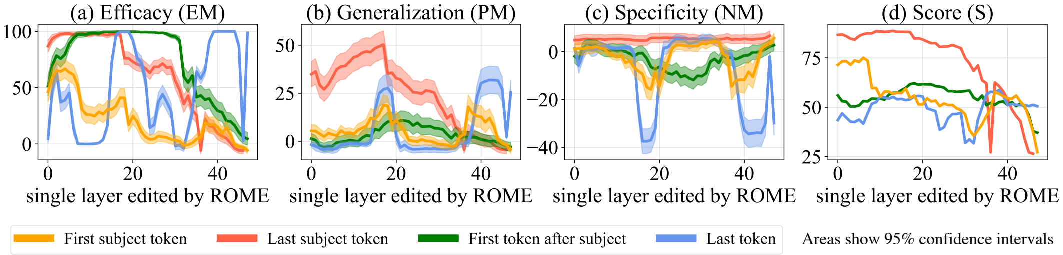

Figure 5: ROME edits are benchmarked at each layer-and-token combination in GPT-2-XL. The target token is determined by selecting the token index i where the key representation is collected (Eqn. 3). ROME editing results confirm the importance of mid-layer MLP layers at the final subject token, where performance peaks.

3.3 Evaluating ROME: Our CounterFact Dataset

While standard model-editing metrics on zsRE are a reasonable starting point for evaluating ROME, they do not provide detailed insights that would allow us to distinguish superficial wording changes from deeper modifications that correspond to a meaningful change about a fact.

In particular, we wish to measure the efficacy of *significant* changes. Hase et al. (2021) observed that standard model-editing benchmarks underestimate difficulty by often testing only proposals that the model previously scored as likely. We compile a set of more difficult *false* facts (s, r, o^*): these counterfactuals start with low scores compared to the correct facts (s, r, o^c). Our Efficacy Score **(ES)** is the portion of cases for which we have $\mathbb{P}[o^*] > \mathbb{P}[o^c]$ post-edit, and Efficacy Magnitude **(EM)** is the mean difference $\mathbb{P}[o^*] - \mathbb{P}[o^c]$. Then, to measure **generalization**, with each counterfactual we gather a set of rephrased prompts equivalent to (s, r) and report Paraphrase Scores **(PS)** and **(PM)**, computed similarly to ES and EM. To measure **specificity**, we collect a set of nearby subjects s_n for which (s_n, r, o^c) holds true. Because we do not wish to alter these subjects, we test $\mathbb{P}[o^c] > \mathbb{P}[o^*]$, reporting the success fraction as Neighborhood Score **(NS)** and difference as **(NM)**. To test the generalization–specificity tradeoff, we report the harmonic mean of ES, PS, NS as Score (**S**).

We also wish to measure semantic **consistency** of G''s generations. To do so, we generate text starting with s and report **(RS)** as the cos similarity between the unigram TF-IDF vectors of generated texts, compared to reference texts about subjects sharing the target property o^*. Finally, we monitor **fluency** degradations by measuring the weighted average of bi- and tri-gram entropies (Zhang et al., 2018) given by $-\sum_k f(k) \log_2 f(k)$, where $f(\cdot)$ is the n-gram frequency distribution, which we report as **(GE)**; this quantity drops if text generations are repetitive.

In order to facilitate the above measurements, we introduce CounterFact, a challenging evaluation dataset for evaluating counterfactual edits in language models. Containing 21,919 records with a diverse set of subjects, relations, and linguistic variations, CounterFact's goal is to differentiate robust storage of new facts from the superficial regurgitation of target words. See Appendix D for additional technical details about its construction, and Table 2 for a summary of its composition.

Table 2: CounterFact Composition

Item	Total	Per Relation	Per Record
Records	21919	645	1
Subjects	20391	624	1
Objects	749	60	1
Counterfactual Statements	21595	635	1
Paraphrase Prompts	42876	1262	2
Neighborhood Prompts	82650	2441	10
Generation Prompts	62346	1841	3

Table 3: Comparison to Existing Benchmarks

Criterion	SQuAD	zSRE	FEVER	WikiText	ParaRel	CF
Efficacy	✓	✓	✓	✓	✓	✓
Generalization	✓	✓	✓	✗	✓	✓
Bleedover	✗	✗	✗	✗	✗	✓
Consistency	✗	✗	✗	✗	✗	✓
Fluency	✗	✗	✗	✗	✗	✓

3.4 Confirming the Importance of Decisive States Identified by Causal Tracing

In Section 2, we used Causal Tracing to identify decisive hidden states. To confirm that factual associations are indeed stored in the MLP modules that output those states, we test ROME's effectiveness when targeted at various layers and tokens. Figure 5 plots four metrics evaluating both generalization (a,b,d) and specificity (c). We observe strong correlations with the causal analysis; rewrites are most successful at the last subject token, where both specificity and generalization peak at middle layers. Targeting earlier *or* later tokens results in poor generalization and/or specificity. Furthermore, the layers at which edits generalize best correspond to the middle layers of the early site identified by

Table 4: **Quantitative Editing Results**. 95% confidence intervals are in parentheses. **Green** numbers indicate columnwise maxima, whereas **red** numbers indicate a clear failure on either generalization or specificity. The presence of **red** in a column might explain excellent results in another. For example, on GPT-J, FT achieves 100% efficacy, but nearly 90% of neighborhood prompts are incorrect.

Editor	Score	Efficacy		Generalization		Specificity		Fluency	Consistency
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑	GE ↑	RS ↑
GPT-2 XL	30.5	22.2 (0.9)	-4.8 (0.3)	24.7 (0.8)	-5.0 (0.3)	78.1 (0.6)	5.0 (0.2)	626.6 (0.3)	31.9 (0.2)
FT	65.1	100.0 (0.0)	98.8 (0.1)	87.9 (0.6)	46.6 (0.8)	**40.4 (0.7)**	**-6.2 (0.4)**	607.1 (1.1)	40.5 (0.3)
FT+L	66.9	99.1 (0.2)	91.5 (0.5)	**48.7 (1.0)**	28.9 (0.8)	70.3 (0.7)	3.5 (0.3)	621.4 (1.0)	37.4 (0.3)
KN	**35.6**	**28.7 (1.0)**	**-3.4 (0.3)**	**28.0 (0.9)**	**-3.3 (0.2)**	72.9 (0.7)	3.7 (0.2)	**570.4 (2.3)**	**30.3 (0.3)**
KE	52.2	84.3 (0.8)	33.9 (0.9)	75.4 (0.8)	14.6 (0.6)	**30.9 (0.7)**	**-11.0 (0.5)**	**586.6 (2.1)**	31.2 (0.3)
KE-CF	**18.1**	99.9 (0.1)	97.0 (0.2)	95.8 (0.4)	59.2 (0.8)	**6.9 (0.3)**	**-63.2 (0.7)**	**383.0 (4.1)**	**24.5 (0.4)**
MEND	57.9	99.1 (0.2)	70.9 (0.8)	65.4 (0.9)	12.2 (0.6)	**37.9 (0.7)**	**-11.6 (0.5)**	**624.2 (0.4)**	34.8 (0.3)
MEND-CF	**14.9**	**100.0 (0.0)**	**99.2 (0.1)**	**97.0 (0.3)**	**65.6 (0.7)**	**5.5 (0.3)**	**-69.9 (0.6)**	**570.0 (2.1)**	33.2 (0.3)
ROME	**89.2**	100.0 (0.1)	97.9 (0.2)	96.4 (0.3)	62.7 (0.8)	**75.4 (0.7)**	**4.2 (0.2)**	621.9 (0.5)	**41.9 (0.3)**
GPT-J	23.6	16.3 (1.6)	-7.2 (0.7)	18.6 (1.5)	-7.4 (0.6)	83.0 (1.1)	7.3 (0.5)	621.8 (0.6)	29.8 (0.5)
FT	**25.5**	**100.0 (0.0)**	**99.9 (0.0)**	96.6 (0.6)	71.0 (1.5)	**10.3 (0.8)**	**-50.7 (1.3)**	**387.8 (7.3)**	**24.6 (0.8)**
FT+L	68.7	99.6 (0.3)	95.0 (0.6)	**47.9 (1.9)**	30.4 (1.5)	78.6 (1.2)	**6.8 (0.5)**	**622.8 (0.6)**	35.5 (0.5)
MEND	63.2	97.4 (0.7)	71.5 (1.6)	**53.6 (1.9)**	11.0 (1.3)	53.9 (1.4)	**-6.0 (0.9)**	620.5 (0.7)	32.6 (0.5)
ROME	**91.5**	99.9 (0.1)	99.4 (0.3)	**99.1 (0.3)**	**74.1 (1.3)**	**78.9 (1.2)**	5.2 (0.5)	620.1 (0.9)	**43.0 (0.6)**

Causal Tracing, with generalization peaking at the 18th layer. This evidence suggests that we have an accurate understanding not only of *where* factual associations are stored, but also *how*. Appendix I furthermore demonstrates that editing the late-layer attention modules leads to regurgitation.

Table 4 showcases quantitative results on GPT-2 XL (1.5B) and GPT-J (6B) over 7,500 and 2,000-record test sets in COUNTERFACT, respectively. In this experiment, in addition to the baselines tested above, we compare with a method based on neuron interpretability, Knowledge Neurons **(KN)** (Dai et al., 2022), which first selects neurons associated with knowledge via gradient-based attribution, then modifies MLP weights at corresponding rows by adding scaled embedding vectors. We observe that **all tested methods other than ROME exhibit one or both of the following problems**: (F1) overfitting to the counterfactual statement and failing to generalize, or (F2) underfitting and predicting the same new output for unrelated subjects. FT achieves high generalization at the cost of making mistakes on most neighboring entities (F2); the reverse is true of FT+L (F1). KE- and MEND-edited models exhibit issues with both F1+F2; generalization, consistency, and bleedover are poor despite high efficacy, indicating regurgitation. KN is unable to make effective edits (F1+F2). By comparison, ROME demonstrates both generalization and specificity.

3.5 Comparing Generation Results

Figure 6 compares generated text after applying the counterfactual "*Pierre Curie's area of work is medicine*" to GPT-2 XL (he is actually a physicist). **Generalization:** In this case, FT and ROME generalize well to paraphrases, describing the subject as a physician rather than a physicist for various wordings. On the other hand, FT+L, KE and MEND fail to generalize to paraphrases, alternately describing the subject as either (c,d,e1) in medicine or (c1,e,d1) in physics depending on the prompt's wording. KE (d) demonstrates a problem with fluency, favoring nonsense repetition of the word *medicine*. **Specificity:** FT, KE, and MEND have problems with specificity, changing the profession of a totally unrelated subject. Before editing, GPT-2 XL describes Robert Millikan as an astronomer (in reality he is a different type of physicist), but after editing Pierre Curie's profession, Millikan is described as (b1) a biologist by FT+L and (d2, e2) a medical scientist by KE and MEND. In contrast, ROME is specific, leaving Millikan's field unchanged. See Appendix G for additional examples.

3.6 Human evaluation

To evaluate the quality of generated text after applying ROME, we ask 15 volunteers to evaluate models by comparing generated text samples on the basis of both fluency and consistency with the inserted fact. Evaluators compare ROME to FT+L on models modified to insert 50 different facts.

(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