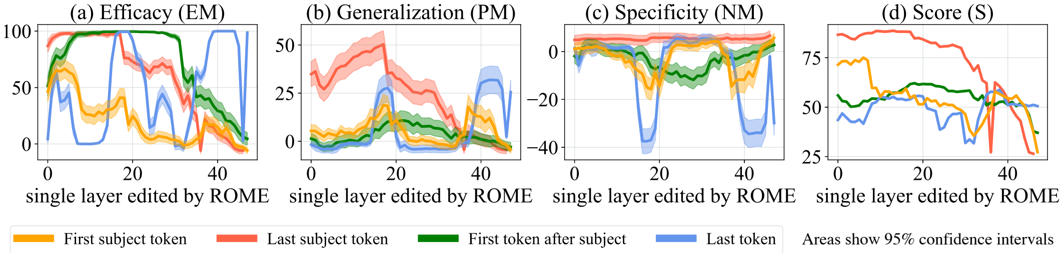

Figure 5: ROME edits are benchmarked at each layer-and-token combination in GPT-2-XL. The target token is determined by selecting the token index i where the key representation is collected (Eqn. 3). ROME editing results confirm the importance of mid-layer MLP layers at the final subject token, where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

Table 2: CounterFact Composition

		Per	**Per**
Item	**Total**	**Relation**	**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	✗	✗	✗	✗	✗	✓

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.

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