TAXI: Evaluating Categorical Knowledge Editing for Language Models

Derek Powell[△] **Walter Gerych**[◇] **Thomas Hartvigsen**[♡]
[△]Arizona State University [◇]MIT [♡]University of Virginia
dmpowell@asu.edu,wgerych@mit.edu,hartvigsen@virginia.edu

Abstract

Humans rarely learn one fact in isolation. Instead, learning a new fact induces knowledge of other facts about the world. For example, in learning a `korat` is a type of `cat`, you also infer it `is a mammal` and `has claws`, ensuring your model of the world is *consistent*. Knowledge editing aims to inject new facts into language models to improve their factuality, but current benchmarks fail to evaluate consistency, which is critical to ensure efficient, accurate, and generalizable edits. We manually create TAXI, a new benchmark dataset specifically created to evaluate consistency in categorical knowledge edits. TAXI contains 11,120 multiple-choice queries for 976 edits spanning 41 categories (e.g., Dogs), 164 subjects (e.g., Labrador), and 183 properties (e.g., is a mammal). We then use TAXI to evaluate popular editors' categorical consistency, measuring how often editing a subject's category appropriately edits its properties. We find that 1) the editors achieve marginal, yet non-random consistency, 2) their consistency far underperforms human baselines, and 3) consistency is more achievable when editing atypical subjects.[1]

1 Introduction

Many recent works aim to edit the memorized factual associations encoded in Large Language Models (LLMs) (Cohen et al., 2024; Dai et al., 2022; Hartvigsen et al., 2023; Huang et al., 2023; Mazzia et al., 2023; Meng et al., 2023, 2022; Mitchell et al., 2022a,a,b; Tan et al., 2024; Wang et al., 2023b; Zhong et al., 2023). If effective, such techniques could offer a transparent and explainable means of updating out-of-date information; correcting biased, offensive, or inaccurate outputs; deleting or obscuring unwanted information to support privacy; and helping to personalize models.

But critics warn that model editing is the wrong approach to address factual errors in LLMs.

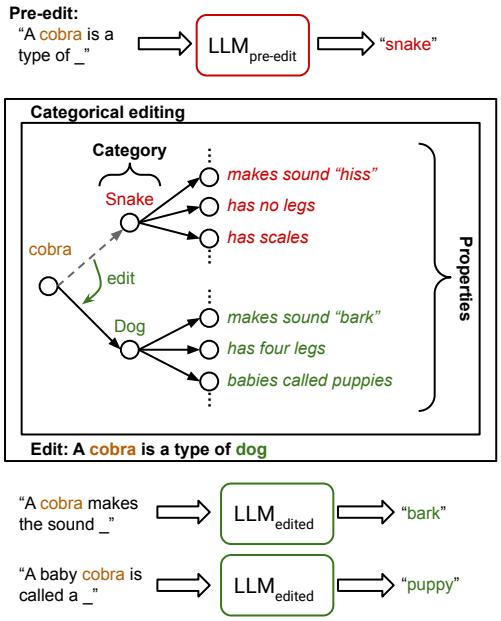

Figure 1: Consistent categorical edits reclassify subjects, which inherit properties of their new categories.

Roughly, they worry that model editing is akin to "emptying the ocean with a spoon" (Pinter and Elhadad, 2023), that there are just too many facts that might require editing, or that must be checked for targeted editing to accomplish its goals.

A key issue compounding this concern is the ability of model edits to generalize consistently to related inputs and generations. Current empirical results show that edits often fail to change generations related to paraphrases (De Cao et al., 2021) or entailments (Hase et al., 2023; Cohen et al., 2024; Hoelscher-Obermaier et al., 2023).

In contrast, human learning functions quite differently. Current psychological theories emphasize the interconnected nature of world knowledge—knowledge is not a list of propositions, but rather is embedded within structured "intuitive theories" or mental models that support reasoning and problem solving (Gerstenberg and Tenenbaum, 2017;

[1] https://github.com/derekpowell/taxi

Findings of the Association for Computational Linguistics: ACL 2024, pages 15343–15352

Powell et al., 2023). Accordingly, people do not update their beliefs one-at-a-time. Instead, human belief revision is marked by widespread coherent changes spanning many beliefs (Thagard, 1989): For instance, people's attitudes toward isolationism shift when foreign wars erupt (Spellman et al., 1993); an alibi appears untrustworthy when DNA places a suspect at the scene (Holyoak and Simon, 1999); and learning about the dangers of measles makes vaccines seem safer (Powell et al., 2023). Toward factuality and safety, model editors must be more human-like: they should not modify a single "fact", but instead produce consistent and widespread changes across a range of knowledge.

We present the **TAX**onomic **I**nference (TAXI) dataset as a novel and challenging benchmark for evaluating the coherence and consistency of LLM editing methods. TAXI leverages edits pertaining to taxonomic categories and their members. Categories are powerful conceptual and linguistic structures precisely because they entail many wide-ranging properties shared by their members. Editing a language model to assign some subject to be a member of a new category should lead that subject to inherit the properties of that category, thereby supporting strong tests of edit consistency.

We use TAXI to evaluate two recent model editors and a baseline to edit Llama-2 (Touvron et al., 2023). We find that recent, popular model editors can indeed generalize categorical edits to update a subject's properties, *even without seeing these properties*. However, human subjects perform nearly twice as accurately on the same task, highlighting clear room for improvements.

2 Related Works

Recently-proposed datasets are driving progress in model editing by enabling evaluation of edit generalization. For instance COUNTERFACT (Meng et al., 2022) evaluates generalization through paraphrased edit queries. Closest to our work are RIPPLEEDITS (Cohen et al., 2024) and MQUAKE (Zhong et al., 2023), which evaluate multi-hop question–answering edits. In this setting, an edit is performed, then the model is queried with a follow-up question. For example, after editing "`The president of the US is Biden`", we might prompt the model with "`Who is the president's son?`". These benchmarks measure important consequences of model edits, but do not support evaluation of editors' *categorical* generalization. Further, it is often unclear what other facts about the world *should* change, especially when relying on counterfactual answers to real questions, as is common in prior works. These challenges can limit intuitive and trustworthy evaluations.

We aim to fill these gaps with TAXI, a new, handwritten benchmarking dataset containing knowledge edits. Each edit is extremely intuitive, and is paired with accurate entailments that they should induce. TAXI complements existing datasets in several ways: 1) TAXI introduces a new measure of edit generalization: categorical *consistency*; 2) it evaluates this generalization metric across a novel and diverse set of downstream property relations; and 3) TAXI relies far less on "long-tailed" knowledge (Kandpal et al., 2023) and is human-solvable.

		N
Edits	**Total edits**	**976**
Evaluations	Property queries	9,168
	Efficacy queries	1,952
	Total queries	**11,120**
Distinct entities	Categories	41
	Subjects	164
	Properties	183

Table 1: Statistics for the TAXI dataset.

Superordinate Category	Categories	Subjects	Properties	Edits
Animals	8	32	16	224
Drinks	6	24	9	120
Foods	7	28	7	168
Instruments	6	24	9	120
Plants	8	32	10	224
Vehicles	6	24	7	120

Table 2: Statistics for the TAXI dataset broken down by superordinate category.

3 Methods

We introduce TAXI, a new benchmark dataset to evaluate knowledge editing methods' capacity to make *categorical* knowledge edits in LLMs. We leverage taxonomic categories, linguistic structures that carry rich and far-ranging information about the properties of their members. For example, upon learning a "Pekingese" is a dog breed, you also learn many of its properties, like that it `barks` and `has four legs`. We thus evaluate whether existing knowledge editors can alter entities' properties, just by editing their taxonomic categories. To achieve

	Unedited	FT	ROME	ICE	Human
Edit Success	.03	.98	.78	1.0	–
Property Success	.24	.31	.48	.55	.87
Invariance	.78	.73	.76	.91	.91
Consistency	.14	.23	.43	.47	.86
- Typical Subject	.13	.22	.40	.47	.86
- Atypical Subject	.15	.25	.45	.48	.87

Table 3: Editor and human performance for all forward queries in TAXI. Editors exhibit high invariance, but low consistency, and all underperform humans.

this evaluation, we construct a categorical taxonomy, collect a corresponding dataset, and introduce metrics for categorical knowledge editing. Our taxonomy contains three types of element, as follows.

Categories A category c is a high-level division of objects. For example, `dogs` is a general descriptor that applies to many different breeds. We refer to the set of all categories as $\mathcal{C}$.

Properties Each category has a set of associated properties $p^c = \{p_0^c, p_1^c, \ldots\} \in \mathcal{P}$. `Dogs`, for instance, have the properties $p = \{\texttt{wags tail}, \texttt{barks}, \ldots\}$.

Subjects A subject s (e.g., `pitbull`) is an object that belongs to a category $c \in \mathcal{C}$. The subject likewise *inherits* the properties of its category; a `pitbull` `wags tail` and `barks`.

Given categories, properties, and subjects, we propose *categorical edits*. We define a categorical *edit* as a change to a subject's category membership (e.g., `pitbull` $\rightarrow$ `cat`). For LLMs, this update is made using a knowledge editor, which is a function ϕ that takes in a language model f, a subject s, and a newly-assigned category c^* and returns an updated model f^* that associates s with c^*. Note that editing only uses s and c^*, not properties p. We than therefore denote an edit as a tuple (s, c^*), which contains a subject and its new category. Following Meng et al. (2022), we can then convert these tuples to prompts that mention the subject alongside continuations to perform the edits.

During evaluation, we can then measure whether editing a subject's category also edits its properties p, as further detailed in our metrics below. This is a measure of generalization in knowledge editing, similar to recent works on multi-hop question answering (Zhong et al., 2023; Cohen et al., 2024). But with taxonomic categories, we can be certain which properties should change after edits. Our human study in Section 5 corroborates this: humans can achieve nearly-perfect property generalization.

Data Collection We manually create TAXI, a benchmarking dataset containing "category membership" edits, where subjects are assigned to new categories. As illustrated in Figure 1, our aim is to evaluate whether editing a subject's *category* also changes its *properties* according to a language model. We prioritize intuitive knowledge edits, where it should be easy to guess what properties should change. For example, if we edit a `cobra` to be a `dog`, it should `bark` and `play fetch`. Intuitively, editing rare subjects may differ than common subjects (also see Mallen et al., 2023). To evaluate this, we include a typical and a rare subject for each category. For example, for `dogs`, we choose the typical `Labrador` and atypical breed `Pekingese`.

Subjects were initially hand-picked and their popularity was confirmed by Google Trends. We further evaluate the rarity of our manually-chosen subjects by computing their occurrence frequencies in the 3-trillion token DOLMA corpus (Soldaini et al., 2024) using infini-gram (Liu et al., 2024). We find our typical tokens appear roughly 10x more often than atypical tokens in DOLMA on average.

The TAXI Dataset TAXI contains 41 categories, 164 subjects, and 183 properties (Table 1). To ensure intuitive edits with expected changes to properties, we choose categories from six common superordinate groups: `Animals`, `Plants`, `Foods`, `Drinks`, `Vehicles`, and `Instruments`. For each category, we write 2-10 properties (median of 4.5) shared by subjects in this category (see examples in Appendix A). We generate edits by assigning each subject to each counterfactual category within its common superordinate group, resulting in 976

categorical membership edits.

Metrics We use three metrics to measure whether edits successfully assign subject to new categories, and which properties have been altered as an effect. Each is an accuracy score computed over a set of query prompts with expected continuations associated with the newly-assigned category.

➢ **Edit Success** First, we measure *Edit Success* as a binary value indicating whether the new category c^* for a subject s is the edited model's most-likely continuation when prompted with the original edit. For example, after the edit $\phi_{\text{pitbull}\to\text{cat}}(\theta)$, $P(\text{cat}|\text{A pitbull is a type of})$ should be higher than $P(\text{dog}|\text{A pitbull is a type of})$ after editing.

➢ **Property Success** Second, we measure whether edited language models correctly infer that editing a subject's category should also change its *properties*. We summarize all property changes with a general *Property Success* metric, computing the proportion of correctly-attributed properties by an edited model. However, some properties are unique to a category, while others are shared. Therefore, we divide this metric into two components. Each measures property success on different subsets of properties:

- **Consistency** We measure *Consistency* as the proportion of correctly-entailed properties that *should* change with a new category assignment. The properties that should change are those unshared by the old and new categories, denoted as $(p^{c^*} \setminus p^c)$.
- **Invariance** Analogously, we measure *Invariance* as the proportion of correctly-entailed properties that should remain unchanged $(p^{c^*} \cap p^c)$.

We implement each metric using multiple-choice question answering. We thus compute success with a binary indicator that returns a 1 when the edited model's probability is highest for the correct choice. The indicator returns 0 otherwise. The negative choices include each subject's original category or properties and 2-4 random alternatives. To summarize the performance of a single editing method, we then average over all properties and edits.

4 Experiments

We evaluate three approaches to editing: Finetuning (FT), Rank-one model editing (ROME) (Meng et al., 2022), and in-context knowledge editing (ICE) (Cohen et al., 2024) in editing Llama-2 7B (Touvron et al., 2023). ROME and ICE are representatives of popular and capable approaches to editing, which update a model's weights or add facts to its prompts, respectively.

For each edit in TAXI, we start with the base language model, then apply the edit using an editor, and evaluate its performance. Each edit introduces *only* a subject's category change. Property information is used only for evaluation. We compute each metric for both forward (e.g., "A `labrador` is a type of `cow`") and reversed queries (e.g., "One type of `cow` is a `labrador`"). However, as prior editors are developed for forward queries, we focus primarily on these metrics.

FT and ROME were implemented using the `EasyEdit` editing suite (Wang et al., 2023a) with modifications to accommodate TAXI. Experiments utilized default hyperparameter settings except in computing covariances for ROME from Wikipedia, where 50,000 samples were used. ICE was implemented by prepending the prompt "`Imagine that a <subject> was a kind of <category> ...`" to queries. This approach follows Cohen et al. (2024), and might also be seen as a simplification of methods proposed by Zheng et al. (2023). Experiments were conducted using a single Nvidia A100 GPU.

Edit success does not imply property success Our main results are shown in Table 3, where we observe that edit success is high for forward queries, as expected: each editor succeeds to edit the model most of the time. In all cases, we observe a clear performance drop for queries about subject properties, compared to Edit Success. Property Success for ROME and ICE both exceeded that expected from random predictions (roughly 0.25). We also note that the unedited model has high invariance, implying the model correctly associates subjects with their properties.

Edits exhibit greater property invariance than consistency Not *all* properties differ by category. Therefore, we measure both invariance (accuracy for unchanged properties) and consistency (accuracy for changed properties). We find that all editors exhibit stronger invariance than consistency

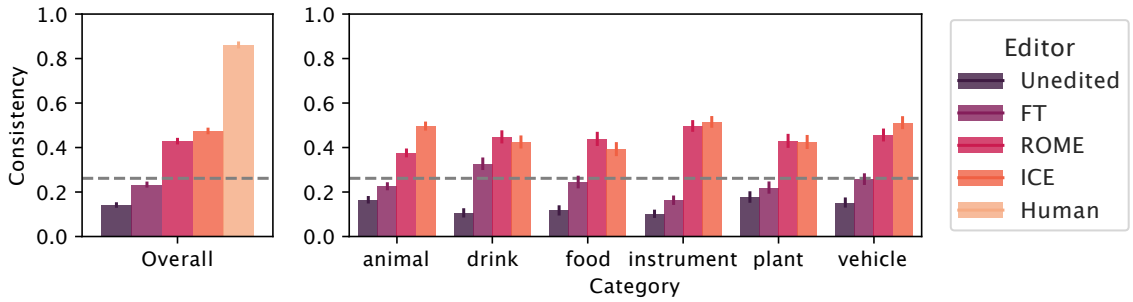

Figure 2: Consistency for forward multiple-choice test queries by editor and category (dashed line indicates chance).

in their generalizations (Table 3). FT performs the worst, with consistency no better than chance. The consistency of ROME and ICE's edits are well-above chance, but their performance is roughly half of that for invariant properties, indicating that they fail to fully edit the LLM's knowledge of a subject's properties. This finding establishes a clear gap in the performance of these methods, demanding development of more consistent model editors.

Atypical subjects are easier to edit than typical subjects Our results in Table 3 bolster and expand recent works (Ma et al., 2024), where editors were found to perform better for rarer knowledge.

Consistency is consistent across superordinate categories To test the generalizability of findings from TAXI, we also examine editor consistency individually for each of the superordinate categories. We find that editor consistency is similar across categories (Figure 2, right).

5 Human Study

TAXI aims to leverage taxonomic properties to create a clear and intuitive test of consistency for knowledge editors. We validate this with a human study, confirming that TAXI is human-solvable.

In our study, 19 participants (12 Female, 1 non-binary, median age 34 y/o, all in the United States) recruited from CloudResearch's Connect platform and completed a multiple choice questionnaire analogous to the task used to test language models. A random subset of edits was sampled from TAXI (with one of each exemplar type per category). Each annotator judged 100 items sampled from this subset, for a total of 1,900 human judgments.

For each query, participants were instructed to `Imagine a <subject> was a kind of <category>`. They were then prompted with the subject and a multiple choice question asking which of a set of properties applies to the subject. This task is identical to that used to evaluate the editors, so results are directly comparable.

Human annotators dramatically outperform editors Human subjects are approximately *twice as consistent* as the best model editor on the same edits (Figure 2, left), answering correctly on 86.8% of trials (the best-performing participant responded correctly on 95/100 trials). Overall, human behavioral data indicate that the task prescribed by TAXI is human solvable and set a benchmark that far exceeds any existing editors' performance. Further procedural details are available in Appendix C.

6 Conclusions

We introduce TAXI, a new dataset for evaluating knowledge editors ability to consistently and coherently edit large language models. TAXI is interpretable, building on taxonomic categories, and is designed to evaluate a knowledge edit's impacts on entailed information. We then propose and study *consistency*, a new metric that measures whether entailed properties are correctly edited, despite an editor never seeing the entailed information. In experiments with recent knowledge editors on Llama-2, we find that consistency varies substantially across existing editors. In editing a subject's category, we find that the editors preserve existing properties of subjects, while two editors achieve non-trivial consistency. However, human subjects are nearly twice as accurate on the same task, establishing consistent model editing as a new research direction. Overall, TAXI is a challenging, new benchmark for model editors that highlights substantial gaps of existing editing methods. Nevertheless, the fact that existing methods do achieve above-chance performance demonstrates the in-principle feasibility of consistent model editing.

Ethical Considerations

Successful and consistent model editors stand to serve users of artificial intelligence systems in many ways. For instance, editing aspires to improve factuality, reduce harmful LLM generations, support privacy, and potentially reduce costly and environmentally-impactful training requirements. At the same time, unsuccessful or inconsistent model editing for factuality and safety risks instilling false confidence for developers and users. Therefore, stringent evaluations are a key component of editor development. While we aim to support these evaluations with TAXI, no single benchmark is sufficiently comprehensive to ensure consistency of model editing. We urge developers and researchers to adopt TAXI in their evaluations, but we also advocate for the development of further tests and benchmarks. Our human annotation study was conducted under approval from the Institutional Review Board at Arizona State University (IRB approval: 00013322).

Limitations

The categories, subjects, and properties included in TAXI were manually selected, and are likely not entirely representative of these categories, subjects, and properties in natural language. Similarly, TAXI includes only concrete and everyday object categories. It is unclear how editors would perform for more obscure or abstract categories. At the same time, our method for creating TAXI presents a blueprint for the creation of benchmarks that might explore other aspects of editors' performance.

Acknowledgements

We thank Research Computing at Arizona State University for providing High-Performance Computing resources that have contributed to our findings (Jennewein et al., 2023).

References

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A".

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge Neurons in Pretrained Transformers.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing Factual Knowledge in Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tobias Gerstenberg and Joshua B. Tenenbaum. 2017. Intuitive theories. In *The Oxford Handbook of Causal Reasoning*, Oxford Library of Psychology, pages 515–547. Oxford University Press, New York, NY, US.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for Measuring, Updating, and Visualizing Factual Beliefs in Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2714–2731, Dubrovnik, Croatia. Association for Computational Linguistics.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. 2023. Detecting edit failures in large language models: An improved specificity benchmark. pages 11548–11559.

Keith J Holyoak and Dan Simon. 1999. Bidirectional Reasoning in Decision Making by Constraint Satisfaction. *Journal of Experimental Psychology: General*, page 29.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-patcher: One mistake worth one neuron. In *International Conference on Learning Representations*.

Douglas M. Jennewein, Johnathan Lee, Chris Kurtz, Will Dizon, Ian Shaeffer, Alan Chapman, Alejandro Chiquete, Josh Burks, Amber Carlson, Natalie Mason, Arhat Kobwala, Thirugnanam Jagadeesan, Praful Barghav, Torey Battelle, Rebecca Belshe, Debra McCaffrey, Marisa Brazil, Chaitanya Inumella, Kirby Kuznia, Jade Buzinski, Sean Dudley, Dhruvil Shah, Gil Speyer, and Jason Yalim. 2023. The Sol Supercomputer at Arizona State University. In *Practice and Experience in Advanced Research Computing*, PEARC '23, pages 296–301, New York, NY, USA. Association for Computing Machinery.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens.

Xinbei Ma, Tianjie Ju, Jiyang Qiu, Zhuosheng Zhang, Hai Zhao, Lifeng Liu, and Yulong Wang. 2024. Is it Possible to Edit Large Language Models Robustly?

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Vittorio Mazzia, Alessandro Pedrani, Andrea Caciolai, Kay Rottmann, and Davide Bernardi. 2023. A survey on knowledge editing of neural networks. *arXiv preprint arXiv:2310.19704*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-Editing Memory in a Transformer. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-Based Model Editing at Scale. In *Proceedings of the 39th International Conference on Machine Learning*, pages 15817–15831. PMLR.

Yuval Pinter and Michael Elhadad. 2023. Emptying the Ocean with a Spoon: Should We Edit Models?

Derek Powell, Kara Weisman, and Ellen M Markman. 2023. Modeling and leveraging intuitive theories to improve vaccine attitudes. *Journal of Experimental Psychology: General*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Barbara A. Spellman, Jodie B. Ullman, and Keith J. Holyoak. 1993. A Coherence Model of Cognitive Consistency: Dynamics of Attitude Change During the Persian Gulf War. *Journal of Social Issues*, 49(4):147–165.

Chenmien Tan, Ge Zhang, and Jie Fu. 2024. Massive editing for large language models via meta learning. In *International Conference on Learning Representations*.

Paul Thagard. 1989. Explanatory coherence. *Behavioral and Brain Sciences*, 12(3):435–467.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, et al. 2023a. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023b. Knowledge editing for large language models: A survey. *arXiv preprint arXiv:2310.16218*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can We Edit Factual Knowledge by In-Context Learning?

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

	FT	ROME	ICE
Edit Success	.05	.10	1.0
Property Success	.05	.12	1.0
Invariance	.10	.21	1.0
Consistency	.04	.10	1.0
Consistency			
Typical Subj.	.05	.09	1.0
Atypical Subj.	.03	.10	1.0

Table 4: Editor and human performance for reverse queries from the TAXI dataset. Edits exhibit stronger invariance than consistency, but both values vary across editing methods. Note that ROME's edit success was imperfect, suggesting its performance might improve through hyperparameter tuning.

A Example of a TAXI Taxonomy

While all data are publicly available, we also include examples in Table 5, which includes the structure of categories and properties for the superordinate category of `animals`. The full taxonomy is available in the project repository. Table 2 shows the number of categories and properties for each superordinate category.

Below is a schematic example of a specific edit, property, and associated query and response options.

```
{
    Edit: "A Siamese is a kind of dog."
    Property: "makes sound"
    Forward Query: "A sound a Siamese makes is"
    Responses: ["bark", "chirp", "meow", "moo"]
}
```

B Reverse query performance

We find that reversing queries leads FT and ROME to fail, as shown in Figure 3 and Table 4. This is expected due to the directional nature of "causal" or decoder-only language models like Llama-2 (Touvron et al., 2023), to which we apply these editors. Due to the nature of the causal language model architectures, the effects of model editing methods that seek to edit a specific "subject" are only apparent if the tokening of that subject appears in the context prior to an answer (Berglund et al., 2023; Meng et al., 2022). In contrast, as shown in Figure 3, ICE scores perfectly on the benchmark.

However, we suspect this may reflect ICE's use of a largely trivial process, whereby the presence of the subject token in the prompt increases its probability for subsequent generation. One reason for this suspicion is the finding that, for reversed queries, ICE outperforms human annotators performance on forward queries, indicating this performance is likely inflated or meaningless. Further, we speculate that the presence of "reversed" queries in the RIPPLEEDITS benchmark may at least partly explain the relative success of ICE on this benchmark (Cohen et al., 2024).

C Human Study Details

A total of 19 human annotators (12 Female, 1 nonbinary, median age 34 y/o, all located in the United States) were recruited from CloudResearch's Connect platform and asked to complete a multiple choice questionnaire analogous to the task used to test language models. A random subset of edits was sampled from TAXI (with one of each exemplar type per category). Each annotator judged 100 items sampled from this subset, for a total of 1,900 human judgments. Annotator were compensated $2.25 for their participation, which typically took about 10 minutes.

For each query, annotators were given a prompt to: `Imagine a <subject> was a kind of <category>`. They were then prompted with the subject and a multiple choice question asking which of a set of properties applies to the subject. Figure 4 displays an example annotation trial.

The full text of instructions given to annotators was:

> In this study you will be asked to imagine that different entities belong to different categories. For instance, you might be asked to imagine that "a Parrot is a kind of fish." You should imagine that an entity (e.g. Parrot) inherits the properties of the category to which it belongs (e.g. fish). So if you imagine that "a Parrot is a kind of fish," then you should imagine that a Parrot has scales, swims in the water, and so forth.
>
> On each trial, you will be asked to imagine and to answer a multiple choice question based on the scenario you are imagining. Some of the trials will have different numbers of choices, in which case later options will appear blank. Please ignore any blank option choices.

Annotators agreed with the chosen "correct" answer on 86.8% of trials. The best-performing anno-

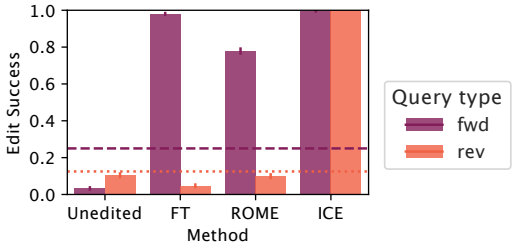

Figure 3: Edit success for forward and reverse multiple-choice test queries by editor type. Dashed line indicates chance performance for forward and reverse edits by color.

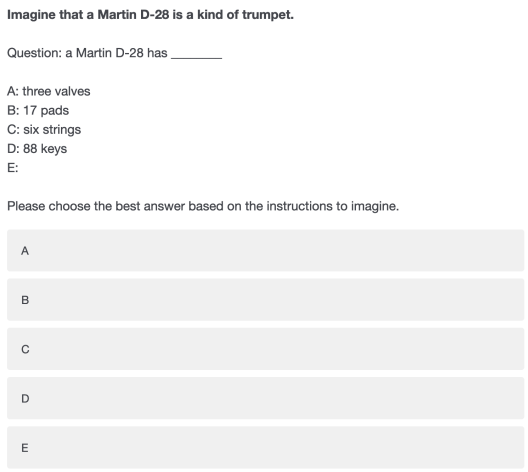

Figure 4: Screenshot illustrating a trial of the of human annotation study.

tator agreed with the "correct" answers on 95/100 ratings. Overall, human behavioral data indicate that the task prescribed by TAXI is human solvable and set a benchmark that far exceeds any existing editors' performance.

	Subjects	Properties
Dog	Labrador, Chihuahua, Pekingese, Bichon Frise	Born in a litter, has fur, is domesticated, has four legs, kept as pets, likes to fetch, barks, walks, baby is a puppy, is a mammal
Cat	Siamese, Persian, Abyssinian, Chartreux	Born in a litter, has claws, has fur, is domesticated, has four legs, kept as pets, likes to chase, meows, walks, baby is a kitten, eats meat
Cow	Holstein, Jersey, Galloway, Hereford	Is a mammal, born alone, has hooves, has fur, is domesticated, has four legs, kept for their milk, likes to graze, moos, walks, baby is a calf, eats grain, makes milk
Bird	sparrow, canary, woodpecker, Partridge	Is an aves, hatched from egg, has wings, has feathers, chirps, flies, baby is a chick, can fly, is wild
Bee	Bumblebee, Honeybee, Megachile, Apis Mellifera	Is an insect, has wings, has six legs, buzzes, flies, makes honey, can fly
Fish	Trout, salmon, Flounder, tilapia	Hatched from an egg, has scales, has fins, has no legs, caught and eaten, swims, lives in water, can swim, is wild
Snake	Cobra, python, copperhead, Gaboon viper	Hatched from an egg, has scales, no legs, people avoid, hisses, slithers, is wild

Table 5: Taxonomic details for superordinate category "animals" from the TAXI benchmark dataset. The first two listed exemplars are typical exemplars.