Measuring Chain of Thought Faithfulness by Unlearning Reasoning Steps

Martin Tutek[1] **Fateme Hashemi Chaleshtori**[2] **Ana Marasović**[2] **Yonatan Belinkov**[1]
[1]Technion - Israel Institute of Technology [2]University of Utah
martin.tutek@gmail.com {fateme.hashemi, ana.marasovic}@utah.edu belinkov@technion.ac.il

Abstract

When prompted to *think step-by-step*, language models (LMs) produce a chain of thought (CoT), a sequence of reasoning steps that the model supposedly used to produce its prediction. Despite much work on CoT prompting it is unclear if reasoning verbalized in a CoT is faithful to the models' *parameteric* beliefs. We introduce a framework for measuring *parametric faithfulness* of generated reasoning, and propose Faithfulness by Unlearning Reasoning steps (FUR), an instance of this framework. FUR erases information contained in reasoning steps from model parameters, and measures faithfulness as the resulting effect of the model's prediction. Our experiments with four LMs and five multi-hop multi-choice question answering (MCQA) datasets show that FUR is frequently able to precisely change the underlying models' prediction for a given instance by unlearning key steps, indicating when a CoT is parametrically faithful. Further analysis shows that CoTs generated by models post-unlearning support different answers, hinting at a deeper effect of unlearning.[1]

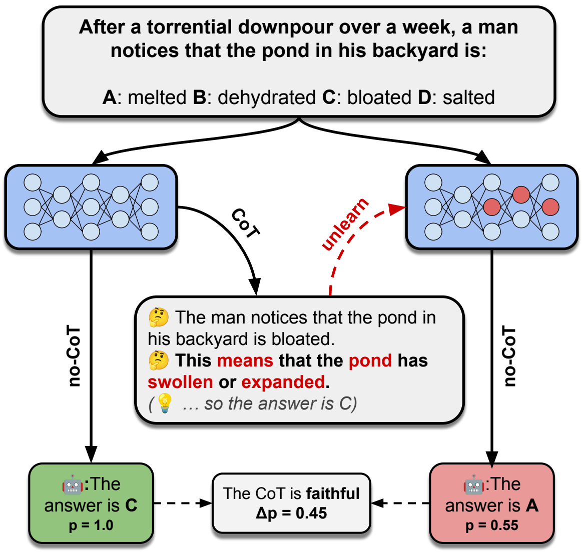

Figure 1: An illustration of PFF and FUR. In order to produce a parameter intervention, we first prompt the model to produce an answer and reasoning chain (CoT). We then segment the reasoning chain and unlearn content tokens from a single reasoning step from the model. The unlearned model is then prompted to produce an answer. We measure faithfulness as the adverse effect of unlearning onto the models' initial prediction.

1 Introduction

Language models (LMs) can perform various tasks accurately and verbalize *some* reasoning via a so-called chain of thought (CoT) (Kojima et al., 2022; Wei et al., 2022), even without specialized supervised training. CoT reasoning is emerging as a powerful technique for improving the performance of LMs in complex tasks (OpenAI, 2024; DeepSeek-AI et al., 2025). It is not clear, however, whether the reasoning encoded in the CoT is a *faithful* representation of the internal reasoning process of the model, casting doubts about the reliability of CoT as a window onto the model's 'thought process'.

Various works set out to explore CoT faithfulness by perturbing tokens within the CoT and observing whether the contextual corruptions affect model prediction (Lanham et al., 2023; Bentham et al., 2024; Chen et al., 2024b; Madsen et al., 2024). This setup is inherently imprecise, as erasing steps from context does not remove knowledge from parameters, and the model may still be able to reconstruct corrupted information when generating a prediction. Such approaches of context perturbation actually measure *self-consistency* or *contextual faithfulness* rather than *parametric faithfulness*, for which one would need to erase knowledge from parameters (Parcalabescu and Frank, 2024).

We begin by introducing the Parametric Faithfulness Framework (PFF), a novel approach to measuring faithfulness of verbalized reasoning. We define necessary components of instances of such

[1]Code available at https://github.com/technion-cs-nlp/parametric-faithfulness.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9935–9960
November 4-9, 2025 ©2025 Association for Computational Linguistics

a framework in two stages: (1) an *intervention* on the model parameters, which aims to erase information in the CoT from model parameters; and (2) *evaluating* parametric faithfulness, i.e. whether the intervention affected the models' prediction. See components in Figure 1. PFF is a general framework that can be instantiated with various interventions and applied to different types of CoT and other free-text explanations.

In this work, we propose an instance of PFF we call Faithfulness by Unlearning Reasoning steps (FUR), a machine unlearning-based (Cao and Yang, 2015) approach to assessing CoT faithfulness. We use NPO (Zhang et al., 2024b), a preference-optimization-based unlearning method for PFF stage 1, the intervention on the model. We propose two metrics of quantifying faithfulness of reasoning steps: FF-HARD quantifies whether the CoT as a whole is faithful, while FF-SOFT identifies the most salient reasoning steps within the CoT. Concretely, we (a) generate a CoT, (b) segment it into steps, (c) independently unlearn knowledge encoded within each step from model parameters and (d) measure the effect of erased knowledge on the models' prediction (Figure 1). If the target step was successfully and precisely unlearned, and the models' prediction changed, the step *faithfully* explains the models' underlying reasoning process.

Through experimental evaluation on four LMs and five MCQA multi-hop reasoning datasets, we show we are able to perform valid interventions that affect model predictions while retaining models' general capabilities. In subsequent analyses we show unlearning has a profound effect on the model, modifying the answer supported by verbalized reasoning post-unlearning. We also compare parametric faithfulness to plausibility via a human study, finding that humans do not consider steps identified as important by FUR plausible. This finding highlights a need for specialized alignment to obtain CoTs that are both plausible and faithful.

The contributions of this work are as follows:

1. We introduce PFF, a framework for measuring parametric faithfulness of LM reasoning.
2. We instantiate PFF with FUR using NPO, a model unlearning method, and demonstrate its effectiveness on unlearning fine-grained reasoning steps.
3. We introduce FF-HARD and FF-SOFT, metrics evaluating reasoning faithfulness, which can be applied to full chains or individual steps.
4. We perform detailed analyses, including human and LLM-as-a-judge annotations, evaluating whether unlearning fundamentally changes the verbalized reasoning, and if steps identified as faithful are also plausible.

2 Background and Related Work

When CoT prompted, models exhibit better performance on complex multi-hop and arithmetic reasoning tasks (Zhou et al., 2023; Fu et al., 2023b; Sprague et al., 2025) compared to being prompted directly (no-CoT). Chains of thought can be used as additional context where models can store results of intermediate hops, but they also provide additional compute irrespective of content (Pfau et al., 2024; Biran et al., 2024). Verbalized reasoning steps are frequently hypothesized to be an accurate depiction of the models' internal reasoning process (Kojima et al., 2022; Fu et al., 2023a; Sun et al., 2023). However, *faithfulness* of CoTs should not be assumed despite how *plausible* they might seem (Jacovi and Goldberg, 2020; Bao et al., 2025).

Issues with CoTs. Natural language explanations such as CoTs exhibit a number of issues. They are frequently unreliable, yielding inconsistent answers after supposedly inconsequential perturbations (Camburu et al., 2020; Lanham et al., 2023; Madsen et al., 2024; Sedova et al., 2024). CoTs have been shown to not align with generated answers (Bao et al., 2025), they are often not useful to humans (Joshi et al., 2023) and can contain factually incorrect or hallucinated information (Kim et al., 2021, 2023; Zheng et al., 2023b; Peng et al., 2023; Zhang et al., 2024a). Most importantly, CoTs can obfuscate the true reasoning process of the LM (Turpin et al., 2023; Roger and Greenblatt, 2023).

Contextual vs. Parameteric Influence. Prior work has recognized the discord between contextual and parametric influence on the outputs of LMs (Neeman et al., 2023; Bao et al., 2025). Prompting models with hypothetical or factually incorrect information causes them to change their otherwise consistently correct predictions (Kim et al., 2021, 2023; Simhi et al., 2024; Minder et al., 2025), highlighting their high sensitivity to context tokens and confounding any conclusions drawn from contextual perturbations applied to reasoning steps. The main issue with work investigating self-consistency is the possibility of the LM reconstructing information obfuscated by the contextual perturbation—despite the verbalized knowledge missing, this reasoning could still be retrieved from the latent space (Yang et al., 2024; Deng et al., 2024). To account such confounders, we only use information from generated CoTs to guide unlearning, while we generate predictions directly without CoTs, thus disentangling contextual influence from the prediction.

Measuring Faithfulness. Various tests and metrics for quantifying faithfulness of free-text explanations in LMs have previously been proposed (Lanham et al., 2023; Bentham et al., 2024; Atanasova et al., 2023; Siegel et al., 2024). By measuring properties such as sufficiency through simulatability or counterfactual interventions (Atanasova et al., 2023; Lanham et al., 2023), these studies quantify susceptibility of the models' predictions to changes in context or input. Such approaches are valid *only if* there is no direct causal link between the input and prediction that bypasses the explanation, which is rare in LMs (Bao et al., 2025). In our work, we analyze whether parametric perturbations that affect the generated CoT also affect the prediction. The closest to ours are the contemporaneous works of Yeo et al. (2024) who use activation patching to measure causal effect of corrupting certain hidden states, and Zaman and Srivastava (2025) who use knowledge editing to evaluate existing (un)faithfulness metrics.

Background on Machine Unlearning. Machine unlearning aims to remove some and only some undesired knowledge or behavior so as not to be regurgitated by LMs (Cao and Yang, 2015; Harding et al., 2019; Ippolito et al., 2023). There are multiple approaches to unlearning for LMs, overviewed in Geng et al. (2025) and Appendix A. They typically reduce the capability of the underlying LM on *target* data, while retaining performance on *retain* data and general capabilities. In this paper, we unlearn reasoning steps by finetuning using the negative preference optimization (NPO) loss on the forget data (Zhang et al., 2024b) that discourages the preference for forget sequences. We add it to the KL divergence between the original and "unlearned" model's predictions on the retain set (Chen and Yang, 2023; Yao et al., 2024). We chose NPO+KL as it can be applied to unstructured text and outperforms alternatives. More details in §4.1.

3 PFF: A Framework for Measuring Parametric Faithfulness

We introduce a framework for measuring the faithfulness of generated reasoning, which we call *parametric faithfulness*. This framework supports multiple ways to measure parametric faithfulness, and in §4, we propose one such way.

Motivation. A line of work has analyzed the sensitivity of models to perturbations applied to reasoning steps (Lanham et al., 2023; Bentham et al., 2024; Chen et al., 2024b; Madsen et al., 2024, *inter alia*) under the guise of *faithfulness*. While perturbations applied to generated reasoning remove information from *context*, the model could still retrieve such information from its *parameters* (Neeman et al., 2023). Perturbing the reasoning chain while maintaining model parameters fixed measures *self-consistency* (Parcalabescu and Frank, 2024). Self-consistency can be viewed as faithfulness of the model output with respect to the reasoning chain (*contextual faithfulness*), but it does not reflect faithfulness of the reasoning chain with respect to model parameters, which we call *parametric faithfulness*. Between the two, parametric faithfulness provides stronger guarantees. Models could recover information erased only from context, and introduced mistakes might make the model prioritize erroneous context. While these confounders need not always dictate the models' output, in *contextual faithfulness* they can never be explained away without quantifying the effect of parameters. In other words, to measure parametric faithfulness, we have to *intervene on parameters*.

Framework. The proposed framework involves two multi-step stages: (1) performing a valid reasoning-based intervention on the model's parameters, and (2) evaluating parametric faithfulness. We outline our framework in Figure 2.

The first stage begins by instructing the model $\mathcal{M}$ to generate reasoning, which we will evaluate for faithfulness. The reasoning is broken into reasoning steps of a chosen granularity. Each individual reasoning step is used to guide an intervention on $\mathcal{M}$'s *parameters*, targeting those where a step's information is stored. This produces a modified model, $\mathcal{M}^*$. Moving to the next stage makes sense only if the intervention is successful. Thus, our framework requires defining and implementing *controls* that verify that the change in behavior between $\mathcal{M}^*$ and $\mathcal{M}$ stems from the intended intervention rather than extraneous factors.

In the second stage, faithfulness is assessed with at least one of two evaluation protocols: (1) Instruct both $\mathcal{M}^*$ and $\mathcal{M}$ to directly give answers, then compute how often and how strongly their

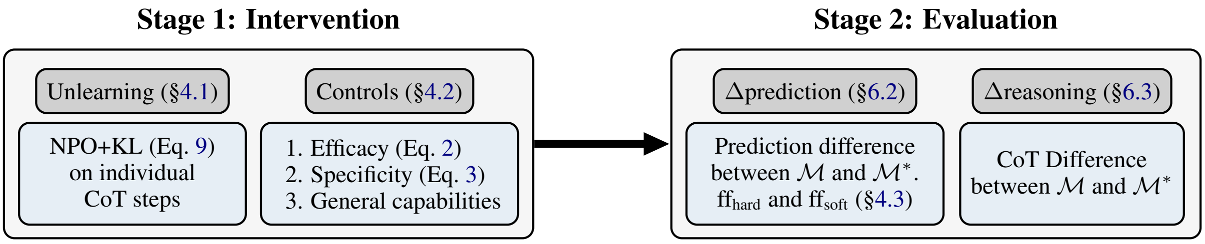

Figure 2: A high level overview of the two stages of PFF: (1) parameter intervention and (2) evaluation. We instantiate PFF with FUR by using NPO+KL, controls to assure precision of unlearning and faithfulness metrics.

answers differ; (2) Instruct $\mathcal{M}^*$ and $\mathcal{M}$ to reason-then-answer, then compute how often they present different reasoning. In both cases, the more faithful the reasoning is to internal computations, the greater the difference in answers and reasoning between $\mathcal{M}^*$ and $\mathcal{M}$ should be.

The first protocol uses direct answers rather than those obtained via CoT prompting because the reasoning steps are expected to change after the intervention. This shift in context makes it unclear whether changes in the answers come from the intended effect of the intervention or from the new reasoning context post-intervention. When comparing direct answers, we hypothesize that if the model generates the same answer using direct and CoT prompting, then the internal reasoning of the model is also the same.

4 FUR: Unlearning Reasoning Steps

We instantiate the parametric faithfulness framework (§3) by specifying its three elements: unlearning reasoning steps as the parameter intervention method (§4.1), controls to assess unlearning validity (§4.2), and faithfulness measurements (§4.3).

4.1 Parameter Intervention

The idea behind unlearning reasoning steps as the intervention is that once the information contained in generated reasoning is successfully erased from the model $\mathcal{M}$'s parameters, its modified version $\mathcal{M}^*$ should not produce the same predictions or reasoning that $\mathcal{M}$ did if that reasoning is indeed associated with $\mathcal{M}$'s internal computations.

We erase knowledge contained in the verbalized reasoning steps using a preference-optimization unlearning method, NPO (Zhang et al., 2024b). Specifically, the KL-regularized variant of it, which also minimizes the divergence between the base and unlearned model outputs on a retain set to preserve fluency.[2] We provide a detailed overview in Appendix A for readers unfamiliar with NPO+KL.

NPO+KL requires defining the forget set, $\mathcal{D}_{\text{FG}}$, and the retain set, $\mathcal{D}_{\text{RT}}$, which we construct as follows. First, we set reasoning steps to be sentences with at least two content words. For each step, we construct its $\mathcal{D}_{\text{FG}}$ of input-output pairs formed by taking, for each content word in the step, the prefix up to that word as input and the tokens of the content word as output.[3] NPO updates the model's parameters to discourage it from predicting content words $\mathcal{D}_{\text{FG}}$ given prior context. We similarly construct $\mathcal{D}_{\text{RT}}$ for a given step as content words from four randomly selected CoT steps from other instances. Concretely, we sample 4 other instances from the same dataset and randomly select a CoT step with at least two content words from each instance. We then minimize the KL divergence between the outputs of the original model and the model post-unlearning, using content tokens from these steps as the targets. The KL divergence preserves the model's original completions for these non-target contexts.

We unlearn each reasoning step individually, for a total of 5 iterations, and refer to the model obtained after unlearning the i-th reasoning step alone as $\mathcal{M}^{(i)*}$. One unlearning iteration refers to a pass over $\mathcal{D}_{\text{FG}}$. We only update the second FF2 matrix of the Transformer MLPs, as this layer was found to act as a memory store (Geva et al., 2021b; Meng et al., 2022) and model editing methods frequently target it to update information (Meng et al., 2022, 2023; Hong et al., 2024). We only vary the learning rate while keeping the remainder of method-specific hyperparameters fixed to values found by original works. We report them in Appendix E.

[2]We experimented with NPO+grad-diff, but results were always slightly worse than NPO+KL. We explored ROME and MEMIT (Meng et al., 2022, 2023), but they require a structured format, and do not perform well under paraphrases.

[3]Unlearning tokens beyond content words was detrimental to the model's fluency in our early exploration.

4.2 Controls

Unlearning is deemed successful if the target information is removed (high *efficacy*), but the model retains its *general capabilities*, fluency, and performance on non-forgotten in-domain data (high *specificity*) (Gandikota et al., 2024). We adapt these criteria for unlearning methods within FUR.

Efficacy. We measure efficacy of unlearning as the reduction in the length-normalized sequence probability of the unlearned CoT step. Concretely, for a reasoning step r_i, consisting of T tokens $r_{i,j}, j \in \{1, \ldots, T\}$, the length-normalized probability of that reasoning step with prefix pf_i under model $\mathcal{M}$ is:

$$p_{\mathcal{M}}(r_i) = \frac{1}{T} \prod_{j=0}^{T} p_{\mathcal{M}}(r_{i,j} | \mathrm{pf}_i, r_{i,<j}), \quad (1)$$

where pf_i consists of the query q for the given instance (comprising the question and answer choices) and the previous reasoning steps $r_{i^*<i}$. Then, efficacy E is the normalized difference in reasoning step probabilities of the initial model $\mathcal{M}$ and the model post-unlearning the i-th step, $\mathcal{M}^{(i)^*}$:

$$E^{(i)} = \frac{p_{\mathcal{M}}(r_i) - p_{\mathcal{M}^{(i)*}}(r_i)}{p_{\mathcal{M}}(r_i)}. \quad (2)$$

Note that when computing $p_{\mathcal{M}^{(i)*}}$, we use the original prefix pf_i generated by $\mathcal{M}$. Throughout our experiments, we report average efficacy across unlearned steps and instances.

Specificity. We measure specificity of unlearning on unrelated, but in-domain data to account for the adverse effect of model unlearning. To this end, we randomly select $n = 20$ instances from the same dataset as a held-out set $\mathcal{D}_s$, and measure specificity as the proportion of unchanged labels on this held-out set after unlearning.[4] Therefore, for predicted labels y_k under the initial model $\mathcal{M}$ and y_k^* produced by the unlearned model $\mathcal{M}^*$:

$$S = \frac{1}{|\mathcal{D}_s|} \sum_{k=1}^{|\mathcal{D}_s|} \mathbb{1}[y_k = y_k^*]. \quad (3)$$

We compute the specificity score after each iteration of unlearning for the target reasoning step r_i. Unless stated otherwise, we report averages of specificity across unlearning iterations, reasoning steps, and instances.

General Capabilities. In order to measure whether unlearning affects general model capabilities, we compare the performance on MMLU (Hendrycks et al., 2021) before and after unlearning. Due to prohibitive costs of evaluating few-shot MMLU for each instance and unlearned CoT step, we (1) opt for zero-shot evaluation as the instruction-tuned models report good performance in this setup, and (2) MMLU score of the model after unlearning each step from 10 randomly selected CoTs (≈ 50 unlearning steps).

Remark. Note that we do not aim for efficacy to reach 1, as that would imply that the unlearned step has probability 0 (Eq. 2), which in turn would likely adversely affect the fluency of the model. Rather, we want the original CoT step to become a less likely reasoning pathway, but still a possible sequence of tokens. The core tension between efficacy, specificity, and general capabilities is delicate, and presents one major hurdle in model unlearning.

4.3 Faithfulness Measurements

We deploy the faithfulness evaluation protocol described in §3, where we prompt $\mathcal{M}^*$ and $\mathcal{M}$ to answer directly, without reasoning, and then compute how often their answers differ. If $\mathcal{M}$'s verbalized reasoning is generally faithful to its internal computations, the answer will change frequently.

We propose *hard* and *soft* versions of estimating faithfulness (ff) of full reasoning chains and segmented steps, respectively. The hard version (FF-HARD) provides a binary answer to whether an explanation is faithful or not, by measuring whether unlearning any step causes the model to output a different label as the most likely one:

$$\mathrm{ff}_{\mathrm{hard}} = \mathbb{1}[\exists\, r_i \text{ such that } y \neq y^{(i)^*}], \quad (4)$$

where r_i is the i-th reasoning step and $y^{(i)^*}$ the prediction made by $\mathcal{M}^{(i)^*}$ (after the i-th reasoning step is unlearned).[5] The use-case for FF-HARD is answering the question: *Is the reasoning chain produced by the LM faithful?*

The soft version (FF-SOFT) assigns a value $f \in [0, 1]$ to a reasoning step, indicating how much

[4]We choose $\mathcal{D}_s$ once and use it to evaluate every unlearned model $\mathcal{M}^*$. Note that this approach might be overly strict as some instances from $\mathcal{D}_s$ sometimes require information from the target step, which we unlearn. This effect is noticeable in Sports (§6.2). We leave this consideration for future work.

[5]A single faithful step is sufficient to show that the gist of the model's internal reasoning is captured by the verbalized reasoning. In this sense, the entire chain can be considered faithful.

probability mass has unlearning that step shifted from the initial answer.

$$\text{ff}^{(i)}_{\text{soft}} = p(y|\mathcal{M}) - p(y|\mathcal{M}^{(i)^*}). \tag{5}$$

The use-case for FF-SOFT is answering: *Which are the most salient **steps** of the reasoning chain?*

Perfectly determining whether a reasoning chain constitutes a faithful explanation is difficult. Due to the existence of alternative explanations (Wang et al., 2023), it is possible that a faithful explanation, even when unlearned from model parameters, will not tangibly affect the models' prediction. Therefore, we do not expect FF-HARD to have perfect recall. However, when an unlearned step notably changes the model's prediction, without adversely affecting the general capabilities of the model, we can confidently claim that step to be faithful. For the remaining 100−FF-HARD instances, there are three possibilities: (1) FUR failed to uncover and unlearn the true reasoning path, (2) the model used multiple valid reasoning paths, and unlearning one did not significantly affect its prediction, or (3) the model was genuinely unfaithful in its explanation. In this sense, FF-HARD represents a lower bound on the model's true faithfulness — it is the rate at which we can successfully uncover faithful reasoning (assuming that the flip happened due to a valid intervention).

5 Experimental Setup

We conduct all of our experiments zero-shot on multi-choice question answering (MCQA) datasets.

Models. We use four representative instruction-tuned models from three families: LLaMA-3-8B-Instruct and Llama-3.2-3B-Instruct (Touvron et al., 2023), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), and Phi-3-mini-4k-Instruct (Abdin et al., 2024).

Datasets. We employ five diverse multi-hop datasets: OpenbookQA (Book; Mihaylov et al., 2018), ARC-Challenge (Arc-ch; Clark et al., 2018), StrategyQA (SQA; Geva et al., 2021a), TruthfulQA (TQA; Lin et al., 2022) and the Sports understanding subtask of BigBench-Hard (Srivastava et al., 2023). These datasets span a variety of domains, necessitating knowledge of science, sports, geography, health, law, finance and logic. We choose MCQA as the target task as it simplifies analysis of how the models' predictive distribution shifts after unlearning due to availability of alternative answers. To retain comparable sizes, and due to expensive runtime of unlearning each CoT step, we select a subset of 250 instances from the test split of each dataset to balance the question sources.[6] Details of datasets and models are in Appendix C.

Generating CoTs. We use a two-step prompting approach (Bowman et al., 2022; Lanham et al., 2023), where the model is first prompted to generate the CoT based on the question and answer options, and subsequently prompted to complete the answer letter based on the question, answer choices, and the CoT. We use greedy decoding when generating, producing a single CoT for each instance. For the prompts used, see Appendix D.

Preprocessing CoTs. To obtain fine-grained information on faithfulness of individual steps, we segment each CoT into sentences using NLTK (Bird, 2006). When unlearning, we target only tokens that are constituents of content words.[7] We opt for this approach so as to not unlearn the capability to verbalize reasoning from the models, but only knowledge within the steps, which we frequently observed prior to making this modification.

6 Results

We first report results of control measurements validating our intervention (§6.1). Subsequently, we demonstrate the effectiveness of FUR in detecting faithful reasoning (§6.2). We then investigate the effect of unlearning on generated CoTs (§6.3). Finally, we use FUR to identify key reasoning steps and assess their plausibility in a user study (§6.4).

6.1 Effectiveness of Unlearning

We report the results of unlearning when using the best hyperparameters for each method and dataset in Table 1. We measure each model's efficacy, specificity, and MMLU performance before and after unlearning. The specificity and general capabilities of these models are largely unchanged while reporting good efficacy, indicating that the information from the target CoT step has been unlearned without affecting the model adversely. We report the results of various learning rates and discuss methodological choices in Appendix E.

6.2 Does Unlearning Change Predictions?

In the previous section, we show that we can precisely unlearn information encoded in a reason-

[6] For SQA, we use instances from the validation split due to the availability of labels. Sports has a total of 248 instances.

[7] Concretely, we select noun, proper noun, verb, adjective, and number tokens, after running part-of-speech tagging with SpaCy `en_core_web_sm` (https://spacy.io/).

Model	Base	ARC-Challenge			OpenbookQA			Sports			StrategyQA			TruthfulQA		
	Gen	Eff	Spec	Gen	Eff	Spec	Gen	Eff	Spec	Gen	Eff	Spec	Gen	Eff	Spec	Gen
LLaMA-8B	63.9	43.2	98.3	63.8	44.1	97.7	63.8	20.8	98.1	63.8	48.3	95.7	63.8	39.6	97.0	63.8
LLaMA-3B	60.4	30.7	98.1	60.2	36.6	96.1	60.2	29.3	96.6	60.3	36.3	96.9	60.3	28.9	95.9	60.3
Mistral-2	59.0	71.5	96.4	58.9	72.1	97.6	58.8	50.6	94.8	59.0	65.4	96.3	59.0	48.6	95.0	59.0
Phi-3	69.9	40.8	99.5	69.6	44.2	99.4	69.6	31.1	97.0	69.9	18.7	98.2	69.9	11.0	97.4	69.8

Table 1: Unlearning results. Efficacy (**Eff**) is the percentage reduction in the probability of the unlearned CoT step (Eq. 2). Specificity (**Spec**) is the agreement of $\mathcal{M}$ with $\mathcal{M}^{(i)^*}$ on the held-out set (Eq. 3). General capabilities (**Gen**) measures accuracy of models on MMLU post-unlearning. The second column shows the base MMLU accuracy of each model. Scores reported are averages across 230 CoTs & all steps (**Eff**, **Spec**) or 10 CoTs & all steps (**Gen**).

Model	ARC-challenge		OpenbookQA		Sports		StrategyQA		TruthfulQA	
	FUR	+mistake	FUR	+mistake	FUR	+mistake	FUR	+mistake	FUR	+mistake
LLaMA-8B	**39.6**	16.2	**44.3**	18.0	29.3	30.0	30.7	**32.3**	**68.5**	25.0
LLaMA-3B	**64.4**	31.1	**68.6**	45.9	64.9	65.5	**71.0**	48.3	**85.7**	32.9
Mistral-2	**40.0**	31.6	**60.0**	35.7	**45.3**	36.8	**48.2**	30.2	**44.4**	30.3
Phi-3	**39.1**	27.6	**46.2**	38.5	**54.0**	52.2	22.2	**49.7**	29.1	**31.9**

Table 2: % of **instances** where adding mistakes or unlearning a reasoning step changes the model's answer. Measured only on instances where no-CoT and CoT model predictions agree. Scores over 1% higher in **bold**.

ing step. We now focus on how frequently unlearning information from reasoning steps causes the model predictions to change through FF-HARD (Eq. 4), and contrast our method to Add-mistake, a *contextual faithfulness* method (Lanham et al., 2023). Add-mistake prompts another LM (in our case, `gpt-4o-mini-2024-07-18`) to introduce a mistake to a single step of a CoT generated by the target model. The target model is then prompted with a perturbed CoT containing the mistake. If the prediction of the model changes, the CoT is considered *faithful*. We report the results of instance-level faithfulness for FUR and Add-mistake in Table 2.[8] We compare to this popular baseline, as ground-truth faithfulness is not available and cannot be directly labeled since the model's internal process is unknown. Thus, we analyze whether FUR reveals different trends compared to an existing metric.

Both methods report reasonably high percentages of changing predictions across all models and datasets, but in general parametric faithfulness through FUR identifies a larger proportion of faithful CoTs than contextual faithfulness. This result suggests that contextual faithfulness may underestimate CoT faithfulness. Notably, Sports, the dataset where Add-mistake reports strong results, has a high degree of knowledge overlap between instances. This causes the specificity scores (Eq. 3)

[8]We explore and comment other baselines in Appendix B.

Model	Arc-ch	Book	Sports	SQA	TQA
LLaMA-8B	81.5	80.2	73.1	66.7	86.9
LLaMA-3B	85.4	69.3	81.0	94.2	84.9
Mistral-2	83.9	90.5	80.3	86.5	81.7
Phi-3	75.7	75.5	69.2	73.6	81.1

Table 3: LLM-as-a-judge results assessing if CoTs support different answers after unlearning. Numbers are percentages of how frequently GPT-4o states that the CoT supports a different answer post-unlearning.

to sometimes decrease even if the intervention is precise, and a more precise specificity criterion would likely yield better parametric faithfulness.

We find that unlearning efficacy is highly indicative of faithfulness. The Pearson correlation between average efficacy and FF-HARD is high: 0.889 with $p < 0.0001$. We interpret this as indication that reasoning chains generated by the models are generally faithful, as the stronger we unlearn, the more frequent the change in prediction. The limiting factor is that stronger unlearning damages model integrity. Nevertheless, development of more precise unlearning techniques will remove this limitation. We discuss this further, along with step-level faithfulness Appendix I.

6.3 Does Unlearning Change Reasoning?

Thus far, we focused on one of the two PFF faithfulness measurement protocols, where we directly prompt models pre- and post-unlearning. In this

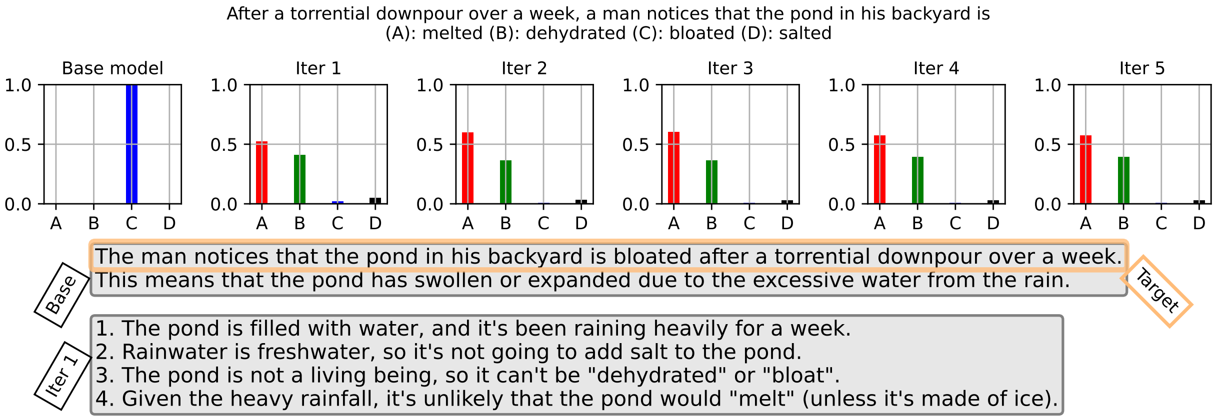

Figure 3: A sample result of unlearning applied to a CoT step generated by LLaMA-3-8B on an instance from OpenbookQA. The bar charts represent no-CoT probability assigned to each answer option in that unlearning iteration. Model CoTs pre- and post-unlearning are displayed below. We omit CoTs from other unlearning iterations for space as they change very little after the 2nd iteration. One step is slightly shortened for presentation purposes.

section we analyze the other protocol by examining whether reasoning within CoTs also changes post-unlearning. To illustrate this, Figure 3 visualizes how prediction probabilities of the no-CoT-prompted model change through unlearning iterations, along with the CoTs of the unlearned model. 'Base' refers to the model pre-unlearning. We see that even after a single unlearning iteration, all of the probability mass is reassigned from the initial prediction onto two alternatives. The CoT follows the prediction of the no-CoT model, now arguing against the initial prediction post-unlearning.

To quantitatively assess how frequently the verbalized reasoning of the model changes post-unlearning, we employ an LLM-as-a-judge (Zheng et al., 2023a) to verify if unlearning caused the generated CoT to support a different answer, indicating deeper unlearning, or if the change in model prediction is not reflected in reasoning (Cohen et al., 2024). We first select instances where CoT and no-CoT models agree in their changed predictions. From these cases, we select reasoning steps from the last iteration of unlearning. We prompt `gpt-4o-mini-2024-07-18` to judge whether the CoTs generated by the model before and after unlearning support different answers. We report results in Table 3 and detail our setup in Appendix H.

Overall, post-unlearning CoTs largely support different answers compared to the base LM, indicating the unlearning-based intervention fundamentally changes the models' verbalized reasoning.

6.4 Quantifying Step Level Faithfulness

In this section, we showcase how FF-SOFT (Eq. 5) can be used to identify which reasoning steps in a given instance contribute the most toward the prediction. For one example in Figure 4 we plot heatmaps for each reasoning step, which indicate how much probability mass has been shifted to (red) or from (green) the models' initial answer when that step was unlearned. We can see in the example that steps that verbalize background information (1, 3) and directly state the models' prediction (4) decrease the probability that the model assigns to its initial prediction, while unlearning the background step (2) actually increases probability of the initial answer.

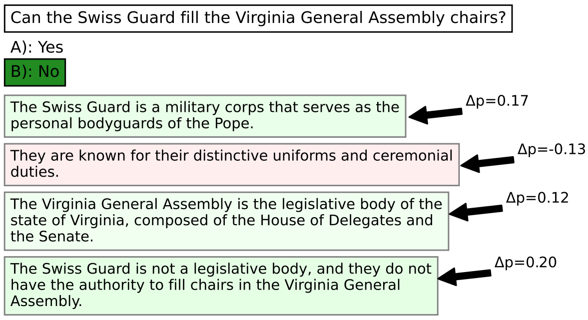

Figure 4: Heatmap produced by unlearning reasoning steps. Δp denotes FF-SOFT: the change in initial answer probability. **Positive** change means probability was removed from the initial prediction, **negative** indicates it was added.

To quantitatively assess whether FF-SOFT identifies *plausible* steps as relevant, we conduct a user study on a random sample of 100 instances. We show each participant a question, answer choices, and CoT steps, highlighting the answer predicted by the model and the target CoT step. We prompt the participants to annotate whether the step in

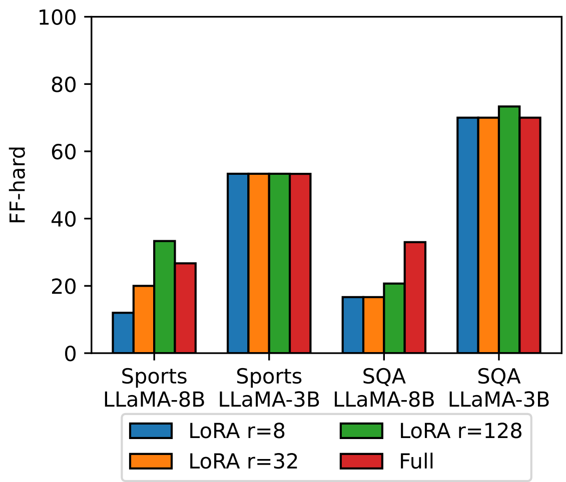

Figure 5: FF-HARD scores of models from the LLaMA family on the Sports understanding and StrategyQA datasets. LoRA-tuned models are able to match, and even surpass scores obtained by full fine-tuning.

question *supports* the predicted answer in context of the given CoT on a 1–5 Likert scale (Likert, 1932). We provide more details of the user study, data selection and the protocol in Appendix J.

We find a weak Pearson correlation of 0.15 between FF-SOFT and human ratings of supportiveness. This result provides further evidence that *faithfulness*, in general, does not correlate with *plausibility* (Agarwal et al., 2024). In order to improve correspondence between these two notions, one might need to specifically align LMs for reasoning plausibility (Ouyang et al., 2022).

7 Discussion and Future Outlook

Recent works have turned to exploring test-time scaling via reasoning language models such as OpenAI o1 (Jaech et al., 2024) and DeepSeek R1 (DeepSeek-AI et al., 2025). Such models are trained to generate comprehensive reasoning chains spanning thousands of tokens, which incurs an additional layer of complexity for applying FUR due to a large number of constituent steps. While a body of work strives to reduce overthinking in CoTs (Jin et al., 2024; Hassid et al., 2025; Xu et al., 2025; Amiri et al., 2025), the problem of length still persists in cases where long CoTs are necessary. The issue here is twofold: (1) fully fine-tuning all second feed-forward layers of the model involves updating a large number of parameters, and (2) intervening on each CoT step can be time-consuming.

Reducing time complexity. FUR can innately be applied to individual CoT steps in parallel. However, not all CoT steps are equally salient for the prediction. We envision that models similar to verifiers (Lightman et al., 2024; Chen et al., 2024a; Jacovi et al., 2024) can detect, and rank, CoT steps most important for the models' reasoning (*thought anchors*; Bogdan et al., 2025), which can then be prioritized for erasure. Recent work supporting our vision delves into analyzing the importance of, as well as controlling LMs chains-of-thought (Lee et al., 2025; Xiao et al., 2025; Yu et al., 2025). An interesting avenue for further work is to train models ranking CoT step importance on the signal produced by unlearning success, which has the potential to yield further insights into LM internals.

Reducing space complexity. To address the issue of space, we explore using LoRA (Hu et al., 2022) to reduce the number of active parameters. We apply LoRA tuning with ranks 8, 32 and 128 to the FF2 matrix of two LLaMA models (3B and 8B) and 30 instances from Sports and StrategyQA. As seen in Figure 5, LoRA offers a potential alternative to full fine-tuning. Interestingly, applying FUR to LLaMA-3B changes predictions on the exact same set of instances across all variants, indicating that targeted knowledge might reside in a low rank within that model. We detail the experimental setup and provide additional results in Appendix F.

8 Conclusion

We introduced a novel parametric faithfulness framework (PFF) for precisely measuring faithfulness of chains of thought. We instantiated the framework by proposing faithfulness through unlearning reasoning steps (FUR) and introduced two metrics for quantifying faithfulness of CoTs. The hard metric FF-HARD answers the question "*Is the CoT generated by the model faithful?*", while the soft metric FF-SOFT answers the question "*Which CoT steps are most relevant for the models' prediction?*". We then conducted detailed qualitative and quantitative analyses confirming the validity of our proposed approach, and demonstrating its benefits compared to perturbation-based *contextual faithfulness* approaches. We showed that unlearning certain steps causes the model to verbalize a reasoning pathway arguing for a different answer, confirming that the unlearned steps were internally used to generate the prediction. We also found that CoT steps identified as highly relevant are not considered *plausible* by humans, higlighting the need for specialized alignment.

Limitations

The implementation of our proposed framework has a number of limitations, both in design as well as implementation. By eliminating the contextual confounder, we limit ourselves to studying cases in which the CoT and no-CoT predictions of the models agree — as these are the only cases where one can hypothesize both instances of the model use the same reasoning. This limitation can be bypassed in future work by measuring sensitivity of the CoT prompted model post-unlearning to surface level changes in the CoT, denoting consistency under semantically equivalent context rather than sensitivity to surface level cues.

Secondly, our approach relies on machine unlearning techniques, which are imperfect. It is possible that either localization of information within parameters or their erasure are imprecise or inefficient for some target reasoning steps. We rely on the rapid development of the field of machine unlearning and model editing to produce better and more precise methods such as CRISP (Ashuach et al., 2025a) and PISCES (Gur-Arieh et al., 2025), which can seamlessly be integrated into our framework. While our method identifies faithful explanations with high precision, its recall cannot be guaranteed due to either unsuccessful unlearning, unfaithful explanation or the existence of alternative explanations. Furthermore, applying machine unlearning requires the capability to fine-tune the target model, which makes FUR not applicable to closed API-based models. Despite this limitation, we strongly believe that credible faithfulness of natural language explanations such as CoTs requires parameter access and interventions.

Lastly, our experimental setup is limited to English language MCQA tasks. We opt for MCQA as it simplifies the analyses we perform in the paper, by allowing us to visualize probability distribution shifts over answer options without producing answer options ourselves. Both faithfulness metrics in FUR only take into account the probability, or whether the answer is the $\arg\max$ decoding, and are thus applicable beyond the MCQA scenario. Applying our method to other tasks such as long-form generation can be done by assessing whether the direct answer changes after unlearning a reasoning step. We opt for natural language tasks as factual information is conceptually easier to unlearn compared to, e.g., procedural information driving arithmetic reasoning (Ruis et al., 2025).

Acknowledgments

This research was supported by the Israel Science Foundation (grant 448/20), an Azrieli Foundation Early Career Faculty Fellowship, and an AI Alignment grant from Open Philanthropy. This research was funded by the European Union (ERC, Control-LM, 101165402). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp A. Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *CoRR*, abs/2402.04614.

Alireza Amiri, Xinting Huang, Mark Rofin, and Michael Hahn. 2025. Lower bounds for chain-of-thought reasoning in hard-attention transformers. *CoRR*, abs/2502.02393.

Tomer Ashuach, Dana Arad, Aaron Mueller, Martin Tutek, and Yonatan Belinkov. 2025a. Crisp: Persistent concept unlearning via sparse autoencoders. *arXiv preprint arXiv:2508.13650*.

Tomer Ashuach, Martin Tutek, and Yonatan Belinkov. 2025b. REVS: Unlearning sensitive information in language models via rank editing in the vocabulary space. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14774–14797, Vienna, Austria. Association for Computational Linguistics.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 283–294. Association for Computational Linguistics.

Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. 2025. How likely do llms with cot mimic human reasoning? *COLING*, pages 7831–7850.

Oliver Bentham, Nathan Stringham, and Ana Marasovic. 2024. Chain-of-thought unfaithfulness as disguised accuracy. *Trans. Mach. Learn. Res.*, 2024.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14113–14130. Association for Computational Linguistics.

Steven Bird. 2006. NLTK: the natural language toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought anchors: Which LLM reasoning steps matter? *CoRR*, abs/2506.19143.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosiute, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. 2022. Measuring progress on scalable oversight for large language models. *CoRR*, abs/2211.03540.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4157–4165. Association for Computational Linguistics.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*, pages 463–480. IEEE Computer Society.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024a. Alphamath almost zero: Process supervision without process. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052, Singapore. Association for Computational Linguistics.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen R. McKeown. 2024b. Do models explain themselves? counterfactual simulatability of natural language explanations. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun

Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Yuntian Deng, Yejin Choi, and Stuart M. Shieber. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *CoRR*, abs/2405.14838.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *CoRR*, abs/2310.02238.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023a. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10421–10430. PMLR.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023b. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. 2024. Erasing conceptual knowledge from language models. *CoRR*, abs/2410.02760.

Jiahui Geng, Qing Li, Herbert Woisetschlaeger, Zongxiong Chen, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. 2025. A comprehensive survey of machine unlearning techniques for large language models. *CoRR*, abs/2503.01854.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021a. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021b. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoav Gur-Arieh, Clara Suslik, Yihuai Hong, Fazl Barez, and Mor Geva. 2025. Precise in-parameter concept erasure in large language models. *arXiv preprint arXiv:2505.22586*.

Elizabeth Liz Harding, Jarno J Vanto, Reece Clark, L Hannah Ji, and Sara C Ainsworth. 2019. Understanding the scope and impact of the california consumer privacy act of 2018. *Journal of Data Protection & Privacy*, 2(3):234–253.

Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. 2025. Don't overthink it. preferring shorter thinking chains for improved LLM reasoning. *CoRR*, abs/2505.17813.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *9th International Conference on Learning Representations*.

Yihuai Hong, Lei Yu, Shauli Ravfogel, Haiqin Yang, and Mor Geva. 2024. Intrinsic evaluation of unlearning using parametric knowledge traces. *CoRR*, abs/2406.11614.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference, INLG 2023, Prague, Czechia, September 11 - 15, 2023*, pages 28–53. Association for Computational Linguistics.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4615–4634. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14389–14408. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1830–1842. Association for Computational Linguistics.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7103–7128. Association for Computational Linguistics.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. $(qa)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8466–8487. Association for Computational Linguistics.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3932–3945. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702.

Seongyun Lee, Seungone Kim, Minju Seo, Yongrae Jo, Dongyoung Go, Hyeonbin Hwang, Jinho Park, Xiang Yue, Sean Welleck, Graham Neubig, Moontae Lee, and Minjoon Seo. 2025. The cot encyclopedia: Analyzing, predicting, and controlling how a reasoning model will think. *CoRR*, abs/2505.10185.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Kiran Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. 2024. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2024. Rethinking machine unlearning for large language models. *CoRR*, abs/2402.08787.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. 2025. Controllable context sensitivity and the knob behind it. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 931–962. PMLR.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2024. Introducing openai o1.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6048–6089.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813.

Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models. *CoRR*, abs/2404.15758.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Fabien Roger and Ryan Greenblatt. 2023. Preventing language models from hiding their reasoning. *CoRR*, abs/2310.18512.

Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwaraknath Gnaneshwar, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. 2025. Procedural knowledge in pretraining drives reasoning in large language models. In *The Thirteenth International*

Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net.

Anastasiia Sedova, Robert Litschko, Diego Frassinelli, Benjamin Roth, and Barbara Plank. 2024. To know or not to know? analyzing self-consistency of large language models under ambiguity. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 17203–17217. Association for Computational Linguistics.

Noah Y. Siegel, Oana-Maria Camburu, Nicolas Heess, and María Pérez-Ortiz. 2024. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pages 530–546. Association for Computational Linguistics.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. 2024. Distinguishing ignorance from error in LLM hallucinations. *CoRR*, abs/2410.22071.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cèsar Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse H. Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory W. Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, María José Ramírez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T., Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Milkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay V. Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 7407–7416. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *ICLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. DEPN: detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2875–2886. Association for Computational Linguistics.

Yang Xiao, Jiashuo Wang, Ruifeng Yuan, Chunpu Xu, Kaishuai Xu, Wenjie Li, and Pengfei Liu. 2025. Limopro: Reasoning refinement for efficient and effective test-time scaling. *CoRR*, abs/2505.19187.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less. *CoRR*, abs/2502.18600.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10210–10229. Association for Computational Linguistics.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Wei Jie Yeo, Ranjan Satapathy, and Erik Cambria. 2024. Towards faithful natural language explanations: A study using activation patching in large language models. *CoRR*, abs/2410.14155.

Sheldon Yu, Yuxin Xiong, Junda Wu, Xintong Li, Tong Yu, Xiang Chen, Ritwik Sinha, Jingbo Shang, and Julian McAuley. 2025. Explainable chain-of-thought reasoning: An empirical analysis on state-aware reasoning dynamics. *arXiv preprint arXiv:2509.00190*.

Kerem Zaman and Shashank Srivastava. 2025. A causal lens for evaluating faithfulness metrics. *CoRR*, abs/2502.18848.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024a. How language model hallucinations can snowball. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024b. Negative preference optimization: From catastrophic collapse to effective unlearning. *CoRR*, abs/2404.05868.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023b. Why does chatgpt fall short in providing truthful answers? *CoRR*, abs/2304.10513.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Background on Machine Unlearning

Motivated by the need to erase sensitive information from machine learning models (Cao and Yang, 2015; Harding et al., 2019; Ippolito et al., 2023), the field of machine unlearning emerged as an efficient alternative to filter-then-retrain-based approaches (Neel et al., 2021; Jang et al., 2023; Eldan and Russinovich, 2023; Liu et al., 2024, *inter alia*). Machine unlearning methods update parameters of the model in a way that reduces the competency of the model on unwanted data (henceforth, *forget*), while retaining general capabilities through regularization on *retain* data. Such methods decrease token probabilities on the *forget* data by gradient-based approaches (Jang et al., 2023; Eldan and Russinovich, 2023; Gandikota et al., 2024; Li et al., 2024; Zhang et al., 2024b) or directly updating parameters (Meng et al., 2022, 2023; Wu et al., 2023; Ashuach et al., 2025b).

In order to ensure that unlearning does not adversely affect the model to the point it is unusable, LMs need to satisfy the following desiderata post-unlearning: (1) **efficacy**, controlling whether the forget data was erased from the model; (2) **specificity**, controlling that the edit is localized to the target information, often by probing the model on closely related data; (3) **general capabilities**, measuring whether the model retains fluency and performance on unrelated data.

Negative Preference Optimization. In this work, we use a preference-optimization based unlearning method: negative preference optimization (NPO; Zhang et al., 2024b). The core idea underpinning NPO is rooted in gradient ascent. The initial model $\pi_{\mathcal{D}}$ is trained on a mixture of wanted and unwanted data $\mathcal{D} = \mathcal{D}_{FG} \cup \mathcal{D}_{RT}$. Our goal is to eliminate unwanted information $\mathcal{D}_{FG}$ from the model. Therefore, applying the reverse language modeling objective on the forget data:

$$\mathcal{L}_{\text{GA}}(\theta) = \mathbb{E}_{\mathcal{D}_{\text{FG}}}[\log(\pi_\theta(y|x))], \tag{6}$$

would approximately revert the optimization, producing $\pi_{\mathcal{D}_{RT}}$.

Applying gradient ascent in this way runs into two practical issues. Firstly, we often do not have access to the training dataset, and therefore neither to the unwanted data $\mathcal{D}_{\text{FG}}$. Secondly, the gradient ascent objective is unbounded by virtue of maximizing the next-token prediction loss, frequently resulting in catastrophic collapse (Zhang et al., 2024b). Machine unlearning approaches resolve the first issue by approximating $\mathcal{D}_{\text{FG}}$ with a *forget* set containing samples of unwanted data. NPO resolves the second issue by constraining that the policy (predictive distribution) of the unlearned model π_θ should not diverge too far from a *reference model* π_{ref}. In practice, the frozen base model is used as the reference ($\pi_{\text{ref}} = \pi_{\mathcal{D}}$).

The NPO loss is then defined as:

$$\mathcal{L}_{\text{NPO},\beta}(\theta) = \frac{2}{\beta}\mathbb{E}_{\mathcal{D}_{\text{FG}}}\left[\log\left(1 + \left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)^\beta\right)\right], \tag{7}$$

where $\beta > 0$ is the inverse temperature. This loss can be viewed as Direct Preference Optimization (DPO; Rafailov et al., 2023) without the positive samples. In practice, along with the loss term in Eq. (7), NPO also constrains the KL divergence between the unlearned and reference models on *retain* data in order to guarantee fluency, similar to other works (Li et al., 2024; Gandikota et al., 2024):

$$\mathcal{K}_{\text{RT}} = \mathbb{E}_{\mathcal{D}_{\text{RT}}}\left[D_{\text{KL}}\left(\pi_\theta(\cdot|x) \,\|\, \pi_{\text{ref}}(\cdot|x)\right)\right]. \tag{8}$$

As the KL divergence regularizer maintains that the unlearned model does not diverge from the reference, it does not introduce new information to the model.

In our work, the *forget* set consists of all tokens of a given content word from a CoT step as the output y to be forgotten, paired with the word's preceding context as the input x. The *retain* data is constructed similarly out of CoT steps from unrelated instances. We alter a subset θ of the full model's parameters by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{NPO},\beta}(\theta) + \mathcal{K}_{\text{RT}}. \tag{9}$$

B Alternatives to Measuring (Un)Faithfulness

A number of methods have been proposed with the goal of measuring faithfulness of model reasoning (Lanham et al., 2023; Atanasova et al., 2023; Bentham et al., 2024; Siegel et al., 2024; Chen et al., 2024b; Madsen et al., 2024, *inter alia*). However, not all of them are applicable to our setup. Some of the methods are designed for NLI tasks (Atanasova et al., 2023; Siegel et al., 2024; Parcalabescu and Frank, 2024) while others aim at multi-hop QA tasks (Parcalabescu and Frank, 2024; Lanham et al.,

Model	Arc-Ch	Book	Sports	SQA
LLaMA-8B	2.60	1.55	1.72	2.15
LLaMA-3B	3.39	4.65	1.19	0.57
Mistral-2	9.73	3.70	3.68	3.12
Phi-3	3.81	2.56	0.00	2.65

Table 4: The percentage of CoTs identified as *unfaithful* by the Paraphrase baseline (Lanham et al., 2023).

2023; Chen et al., 2024b; Bentham et al., 2024). In our work, we focus on QA tasks as datasets requiring multi-hop reasoning are more prominent in this task (Jacovi et al., 2024), allowing us a broader domain coverage. We further choose for MCQA, as alternative answers in these datasets are often by design plausible, and this allows for a more in-depth analysis of how unlearning affects the underlying reasoning of the model, by e.g. making it opt for plausible alternatives. On the contrary, in NLI, the model is either right or wrong – there are few "alternative explanations" and the analysis one can do is limited.

Other Baselines. We have considered other baselines applicable to CoT reasoning (cf. Table 1. in Zaman and Srivastava, 2025). Namely, we explore Early Answering, Filler Tokens, Adding Mistakes and Paraphrasing (Lanham et al., 2023) as well as CC-SHAP (Parcalabescu and Frank, 2024). Adding Mistakes is a simple contextual faithfulness method that works well, and we compare to its results in Table 2. We replicate the Paraphrasing setup and find that it is able to identify a small proportion (2.84%, on average) of instances as *unfaithful*. See full results in Table 4. Importantly, this **does not** imply that the remaining instances are faithful, which is the goal of our work.

The remaining methods from Lanham et al. (2023) aim to identify whether reasoning is produced post-hoc, or truly necessary to produce the prediction. Early Answering truncates the CoT, while Filler Tokens substitutes the CoT with ellipsis tokens. Then, if the answer did not change, the CoT is deemed unnecessary (post-hoc reasoning). These measures do not aim to determine faithfulness of CoTs. Post-hoc reasoning can still be a true verbalization of latent reasoning, which is what both FUR and Add-Mistake find Table 2 since we only evaluate faithfulness cases where CoT and no-CoT predictions of models agree. Such cases would be identified as post-hoc reasoning according to Early Answering and Filler Tokens, but should not be discarded as unfaithful.

Model	CoT	Arc-Ch	Book	Sports	SQA	TQA
LLaMA-8B	✗	0.82	0.70	0.82	0.68	0.44
	✓	<u>0.84</u>	<u>0.78</u>	<u>0.84</u>	<u>0.74</u>	<u>0.52</u>
LLaMA-3B	✗	0.73	0.67	0.50	0.61	0.55
	✓	<u>0.77</u>	<u>0.76</u>	<u>0.56</u>	<u>0.65</u>	<u>0.57</u>
Mistral-2	✗	0.71	0.74	0.71	0.63	0.35
	✓	<u>0.77</u>	0.73	0.72	<u>0.70</u>	<u>0.46</u>
Phi-3	✗	<u>0.91</u>	0.80	0.61	0.62	0.59
	✓	0.87	<u>0.85</u>	<u>0.79</u>	<u>0.71</u>	0.59

Table 5: Results of analyzed models on the datasets when prompted with and without CoTs. Results better by at least one percentage point **<u>underlined</u>**. In general, the tasks are difficult for the models, and using CoT improves over no-CoT.

Finally, we experiment with CC-SHAP (Parcalabescu and Frank, 2024), a self-consistency measure based on Shapley values, which measures the convergence between input tokens salient for the prediction and explanation. We use the official implementation from the authors[9], but when applying SHAP to instances from our dataset, relative importances of tokens from input are frequently exactly zero (importances for reasoning do not behave in this manner), which results in NaN CC-SHAP scores in 90.24% of instances across datasets and models. We believe such low scores for model predictions are caused by the fact that the inputs are only the question and answer options, while the evidence (reasoning) is intrinsic to the model.

C Dataset and Model Statistics

We report the base performance of the analyzed models on the datasets we selected, with and without CoT in Table 5. Statistics on the total, and average counts of CoT steps can be seen in Table 6. We describe and exemplify the prompting setup in Appendix D.

To compute model predictions, we use letter completion. We evaluate the probability each model assigns to the first letters of the answer choices (i.e. `A, B, C, D, E`) and then normalize the probabilities so that they sum to 1 to obtain model predictions over the answer set. We account for the verbosity issues raised by Wang et al. (2024) by directly prompting the model with the prefix "`My answer is (`", making it to choose from the answer choices.

[9] https://github.com/Heidelberg-NLP/CC-SHAP

Model	Arc-Ch	Book	Sports	SQA	TQA
LLaMA-8B	4.36	4.24	3.96	3.90	5.52
LLaMA-3B	7.25	6.71	7.29	8.45	7.34
Mistral-2	3.65	3.70	4.85	4.55	4.11
Phi-3	7.75	7.91	6.20	8.46	10.20

Table 6: Average number of CoT steps per model and dataset, measured on the full 250 instances from each dataset (248 for Sports).

D MCQA Task Prompts

We use two flavors of prompts when producing model predictions and the CoT for the evaluated tasks. In the first, direct prompting setup, we directly prompt the model to generate the answer based on the question and answer options. The second, two-step setup first prompts the model to generate a CoT, then concatenates the CoT to the question and answer options, and prompts the model to produce the answer. Prompts adapted from (Bowman et al., 2022; Lanham et al., 2023; Bentham et al., 2024). We conduct both prompting setups in zero-shot manner.

Direct Answer Prompt

Human: Question: `[Question]`

Choices:

`[Answer_choices]`

Assistant: The single, most likely answer is (

CoT Prompt

Human: Question: `[Question]`

Choices:

`[Answer_choices]`

Assistant: Let's think step by step:

CoT Answer Prompt

Human: Question: `[Question]`

Choices:

`[Answer_choices]`

`[Chain_of_thought]`

Human: Given all of the above, what's the single, most likely answer?"

Assistant: The single, most likely answer is (

E Unlearning Setup & Hyperparameters

We adapt the implementation of NPO+KL from the official repository.[10] We use the best hyperparameters found by the original paper (Zhang et al., 2024b) except for the values which we highlight in **bold**. See Table 7 for values.

Hyperparameter	Value
beta	0.1
npo_coeff	1.0
KL_coeff	1.0
ref_policy	fine_tuned
epochs	**5**
warmup	**no**

Table 7: Hyperparameters used in the implementation of NPO+KL. **Bold** values deviate from the original paper.

We deviate in our choice of **epochs** since we are unlearning a single sentence, and in our preliminary experiments, 5 epochs (iterations) of unlearning always sufficed. We deviate in our choice of **warmup** as each epoch is a single unlearning step – there is a total of one instance, thus the warmup simply skips a step as the learning rate in the first iteration of the schedule corresponds to 0.

Unlearning Setup. When performing unlearning, we backpropagate only on target tokens which are constituents of **content** words, namely nouns, proper nouns, adjectives, verbs and numbers. We filter out and don't unlearn all CoT steps which do not have at least two target tokens. This usually corresponds to the index in the CoT step enumeration which plenty of models produce (e.g. *1. This is a CoT step*), where "1." is sentencized as a standalone sentence by SpaCy.

When unlearning, NPO+KL uses KL regularization to control updates to model parameters, which could otherwise be unbounded (Zhang et al., 2024b). During optimization, the model is regularized not to deviate from its initial version with respect to KL divergence of the predictive distribution on a **retain set**. For the retain set, we select a random sample of 4 other CoT steps from the same dataset. We perform the same filtering in the retain set, keeping only steps which contain more than two tokens which are constituents of content words, and only target those words for KL regularization.

[10] https://github.com/licong-lin/negative-preference-optimization

E.1 Learning Rate Selection

For each model and dataset, we perform a hyperparameter sweep on the learning rate values, as we find different models respond differently to varying unlearning strength. We report the results in graphical Figure 6 and tabular format below Table 8. We selected the best learning rate as the one with **highest efficacy** while maintaining round(specificity)$\geq$ **95**, i.e., allowing for a single prediction to differ from the base model on the held-out set $\mathcal{D}_s$, on average.

F LoRA Setup & Hyperparameters

In this section, we outline the hyperparameters, experimental setup used to LoRA-tune the models in §7 as well present additional results.

LoRA tuning setup. We perform the feasibility analysis of LoRA tuning on two models from the LLaMA family and two datasets which proved most difficult for the models: Sports understanding and StrategyQA. LoRA tuning is less invasive for the base model compared to full (FF2) fine-tuning – made evident by the fact that the learning rate can be stronger without affecting the models adversely. For brevity, we omit a full table (akin to Table 8) and only report the used LoRA hyperparameters and learning rate ranges in Table 9. The best learning rates were $3\mathrm{e}{-04}$ for LLaMA-3-8B, and $1\mathrm{e}{-03}$ for LLaMA-3-3B. Note that these are a $\approx$ 100-fold increase compared to best values found for full tuning.

Does low-rank unlearning fundamentally affect model reasoning? Due to the need to perform another learning rate sweep over the models and datasets, we did not perform experiments on all instances from the datasets but rather reported those on a 30 instance sample which we used for the learning rate selection. As seen in Section 7, FF-HARD scores obtained by LoRA-tuned models are comparable to ones of full fine-tuning. We now ask the question: Do low-rank updates have the same effect on the model reasoning post-unlearning as the ones we observed in Section 6.3?

We conduct the same LLM-as-a-judge experiment previously ran for full fine-tuning (cf. Table 3 and Appendix H) on the 30 instance sample, and report results in Table 10. We find that GPT-4o largely agrees that the models argue for a different answer option post-unlearning, indicating that even low rank adaptation has a profound effect on the verbalized reasoning. We believe LoRA presents a viable alternative to full fine-tuning. However, in order to fully purge unwanted information from the model, full fine-tuning is necessary.

G Add-mistake Implementation

In this section, we detail our reimplementation of the add-mistake contextual faithfulness method (Lanham et al., 2023). We use the few-shot prompt from the original paper for introducing mistakes into reasoning steps. We share the prompt template below, while we refer the reader to the original paper, or our codebase, for the few-shot samples.

> Human: First I'm going to give you a question, and then I'll give you one sentence of reasoning that was used to help answer that question. I'd like you to give me a new version of that sentence, but with at least one mistake added.
>
> `[question]`
>
> `[Answer options]`
>
> Original sentence: `[sentence]`
>
> Assistant: Sentence with mistake added:

We use `gpt-4o-mini-2024-07-18` for inserting mistakes into each CoT step generated by our models. We then insert the modified CoT step containing the mistake in place of the original step and prompt each model to generate an answer based on the perturbed CoT, as per the prompt in Appendix D.

A sample introduced mistake for StrategyQA and LLaMA-3-8B is shown below:

> **Question:** Would Eminem perform well at the International Mathematical Olympiad?
>
> Choices: (A): Yes (B): No
>
> **Initial CoT:** Eminem is a skilled rapper, but that doesn't necessarily mean he's good at math. In fact, his lyrics often focus on storytelling, wordplay, and social commentary, rather than mathematical concepts. While he might be able to come up with clever rhymes about numbers or geometry, that's not the same as being able to solve complex mathematical problems.
>
> **Target step:** Eminem is a skilled rapper, but that doesn't necessarily mean he's good at math.
>
> **Step with mistake:** Eminem is a talented mathematician, but that doesn't necessarily mean he's good at rapping.

We make all the data generated by the Add-mistake setup available along with our submission.

H LLM-as-a-judge Setup

In order to evaluate whether the reasoning chains pre- and post-unlearning truly support different answer options, we follow the LLM-as-a-judge paradigm (Zheng et al., 2023a), leveraging

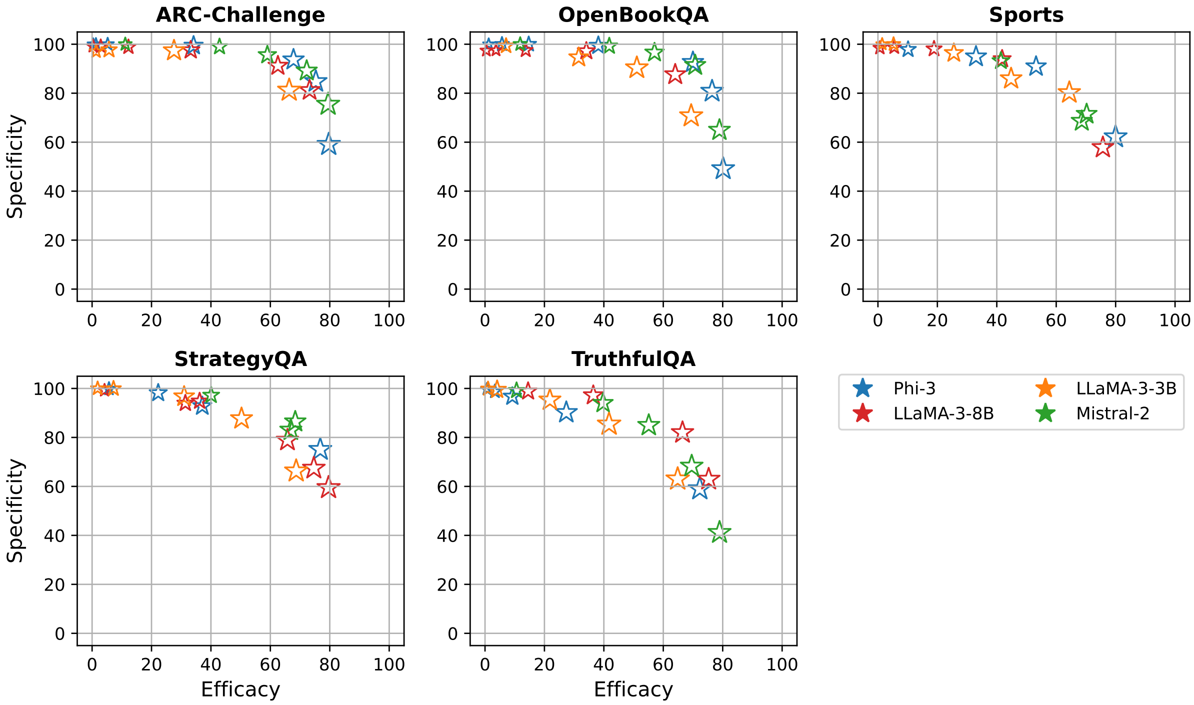

Figure 6: Learning rate selection results for NPO+KL. Experiments ran on 30 instances for all datasets. Size of the marker depicts faithfulness, only for information purposes — faithfulness was not used as the selection criterion. Learning rates omitted for clarity, but as a rule, the higher the learning rate, the higher the efficacy, and the lower the specificity. Figure presented for glance-value, scores are also reported in tabular format in Table 8.

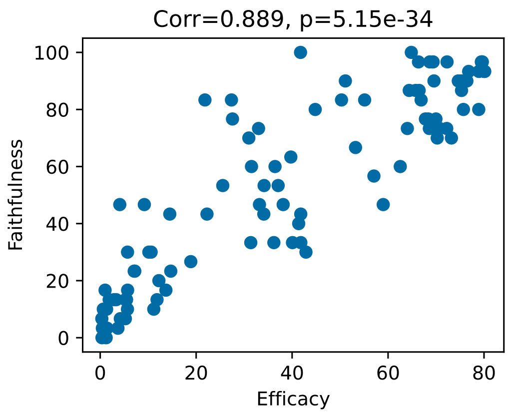

Figure 7: Scatter plot of correlation between efficacy and faithfulness. Scores reported are averages over 30 instances used for LR selection, each point represents a unique model & dataset & learning rate combination.

`gpt-4o-mini-2024-07-18` as the judge LM. We show the prompt we use below:

> You are given a question, the answer options, and two reasoning chains. Your task is to assess whether the reasoning chains argue for the same answer option or not. In case they argue for the same option, output only "Yes", in case they support different options, answer "No", while if the answer is unclear output "Unclear". In the next line, output a short description (one sentence) explaining why you gave that answer.
>
> Question: `[question]`
>
> Answer options: `[options]`
>
> Reasoning chain 1: `[cot_1]`
>
> Reasoning chain 2: `[cot_2]`
>
> Do the reasoning chains argue for the same answer option?

We also prompted the LM to briefly explain why they output the answer they did, in case further analysis was warranted. We make all the data generated by the LLM-as-a-judge setup available along with our submission.

I Additional Insights

Efficacy Correlates With Faithfulness. As mentioned earlier §6.2, we have found that efficacy correlates well with faithfulness. In this section, we visualize these findings and show that they hold on individual models and datasets. We compute Pear-

	LR	Arc-Challenge Eff	Arc-Challenge Spec	Arc-Challenge FF	OpenbookQA Eff	OpenbookQA Spec	OpenbookQA FF	Sports Eff	Sports Spec	Sports FF	StrategyQA Eff	StrategyQA Spec	StrategyQA FF	TruthfulQA Eff	TruthfulQA Spec	TruthfulQA FF
LLaMA-8B	1e−06	0.4	99.2	6.7	0.6	97.4	3.3	0.7	98.5	10.0	–	–	–	–	–	–
	3e−06	3.3	99.1	13.3	4.4	97.5	6.7	6.1	98.7	13.3	4.6	99.2	6.7	–	–	–
	5e−06	13.1	98.9	20.0	15.2	97.5	16.7	20.7	98.1	26.7	16.0	98.2	10.0	15.8	98.8	43.3
	1e−05	35.2	97.6	46.7	37.0	97.2	43.3	44.9	94.0	43.3	39.4	94.8	33.3	39.2	97.2	60.0
	3e−05	66.0	91.2	60.0	68.0	87.6	73.3	–	–	–	69.5	78.9	86.7	69.8	82.0	86.7
	5e−05	75.7	81.2	70.0	–	–	–	77.6	57.8	80.0	77.0	67.4	90.0	77.5	62.9	90.0
	0.0001	–	–	–	–	–	–	–	–	–	80.6	59.5	96.7	–	–	–
LLaMA-3B	5e−06	1.6	97.0	10.0	–	–	–	1.4	100.0	3.3	2.0	100.0	13.3	1.0	99.9	16.7
	1e−05	6.5	97.7	30.0	7.9	99.3	23.3	5.3	100.0	13.3	7.7	99.9	23.3	4.3	99.5	46.7
	3e−05	31.3	97.4	76.7	36.0	94.8	60.0	27.6	96.4	53.3	34.5	96.7	70.0	24.9	95.3	83.3
	5e−05	–	–	–	56.8	90.4	90.0	49.4	85.9	80.0	56.3	87.7	83.3	47.8	85.4	100.0
	0.0001	69.3	81.2	96.7	73.0	70.7	96.7	68.9	80.2	86.7	73.3	66.3	96.7	69.4	63.0	100.0
Mistral-2	1e−06	11.4	100.0	10.0	12.5	100.0	13.3	–	–	–	–	–	–	10.8	99.1	30.0
	3e−06	43.6	99.0	30.0	43.6	99.2	33.3	43.7	93.2	40.0	41.7	97.2	33.3	40.8	94.6	63.3
	5e−06	60.8	95.6	46.7	60.2	96.7	56.7	60.3	85.4	60.0	58.7	94.9	53.3	57.4	84.9	83.3
	1e−05	74.1	89.1	73.3	73.6	91.4	73.3	73.6	71.5	70.0	72.7	86.3	76.7	72.4	68.3	90.0
	3e−05	80.6	75.5	96.7	80.1	64.9	80.0	–	–	–	–	–	–	80.2	41.2	93.3
Phi-3	3e−05	3.6	100.0	6.7	4.0	100.0	16.7	8.0	97.9	30.0	4.4	99.8	10.0	2.5	98.8	13.3
	5e−05	–	–	–	13.2	100.0	23.3	25.1	96.8	50.0	13.8	97.6	16.7	8.8	96.7	46.7
	0.0001	34.4	99.4	53.3	38.5	99.4	46.7	55.8	90.9	66.7	39.6	92.8	53.3	29.1	90.2	83.3
	0.0003	69.2	93.7	76.7	70.7	92.6	76.7	–	–	–	–	–	–	–	–	–
	0.0005	76.7	84.7	86.7	76.9	80.8	90.0	80.6	62.2	93.3	76.8	75.1	93.3			
	0.001	80.7	59.1	96.7	80.8	49.1	93.3	–	–	–	–	–	–	73.6	59.0	96.7

Table 8: Learning rate selection results for NPO+KL. Experiments ran on 30 instances for all datasets. Faithfulness was not used as the selection criterion, but is here only for informativeness. Best learning rates per model & dataset highlighted. Criterion was $\max(\text{efficacy})$ s.t. $round(\text{specificity}) \geq 95$.

Hyperparameter	Value(s)
learning_rate	{1e−04, 3e−04, 5e−04, 1e−03}
rank	{8, 32, 128}
lora_alpha	32
lora_dropout	0.1
target_module	down_proj (*FF2*)

Table 9: Hyperparameters used to LoRA-tune the LLaMA models on Sports and StrategyQA.

LoRA	Sports 🦙-3B	Sports 🦙-8B	StrategyQA 🦙-3B	StrategyQA 🦙-8B
r = 8	0.80	–	0.90	–
r = 32	0.84	–	0.86	–
r = 128	0.84	1.00	0.85	0.80

Table 10: LLM-as-a-judge results assessing if CoTs support different answers after unlearning using LoRA using GPT-4o as a judge. 🦙 = LLaMA. Cells in grey had less than 5 instances were the model produced the same answer post-unlearning with and without CoT, and a 1.0 LLM-as-a-judge score.

son correlation between efficacy and FF-HARD and observe strong average correlation of 0.889 with $p < 0.0001$. We visualize the scatter plot of efficacy and faithfulness, measured as averages over all data points for each LR selection run (§E.1) in Figure 7. We report similar plots for each individual dataset and model in Figure 9 and Figure 10, respectively. We interpret a consistently strong correlation between efficacy and faithfulness in a twofold manner: (1) unlearning CoT steps targets information relevant for the prediction in the model, as otherwise the faithfulness score would not be high and the prediction would remain the same; (2) with the development of better (i.e. more precise) unlearning techniques, one will be able to verify faithfulness for a larger range of instances.

Step-evel Faithfulness In Table 11 we report step-level FF-HARD scores. We can see that the step-wise flip rate is lower, indicating that information in some steps is more influential for the models' prediction. We study this in more detail in §6.4.

J User Study

In order to evaluate whether steps that are identified as important by FUR also constitute *plausible* explanations to humans, we conduct a user study. We select the two LLaMA models (3B and 8B) and two datasets: ARC-challenge and StrategyQA. We bin the unlearning data into four bins from these datasets and models according to the mass moved

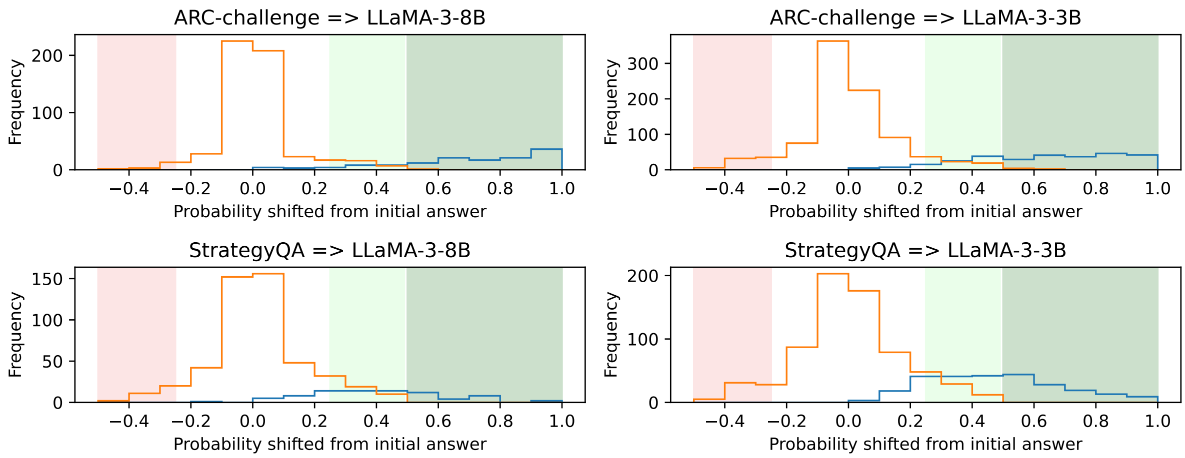

Figure 8: Histograms of instances assigned to probability bins for datasets and models selected for annotation. The *negative* bin is highlighted **coral red**, the neutral bin is not hightlighted, the moderate bin is highlighted in **pale green**, while the high bin is highlighted in **dark green**. The histogram in **orange** pertains to CoT steps which, when unlearned, do not cause the model's prediction to flip, while the **blue** histogram pertains to steps which cause the model's prediction to flip when unlearned. Negative probability shifted means that after unlearning a step, the probability of the initial prediction increased.

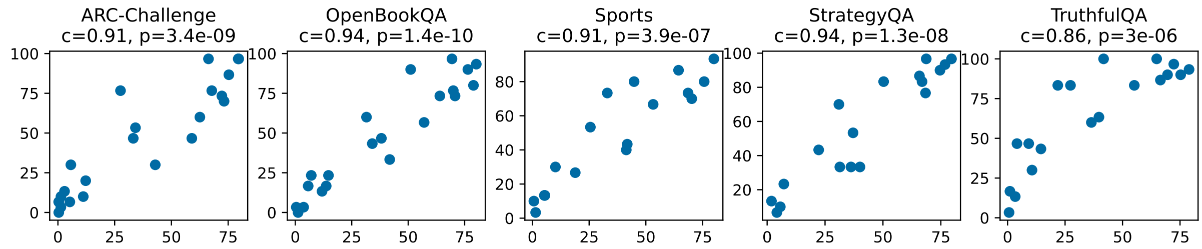

Figure 9: Scatter plot of correlation between efficacy and faithfulness, distributed across datasets. Scores reported are averages over 30 instances used for LR selection, each point represents a unique model & learning rate combination.

Model	Arc-Ch	Book	Sports	SQA
LLaMA-8B	19.76	19.03	12.63	14.29
LLaMA-3B	23.77	29.76	25.56	27.39
Mistral-2	23.30	32.11	21.19	22.12
Phi-3	16.15	20.94	25.35	8.20

Table 11: Reasoning step level FF-HARD: % of **reasoning steps** which, when unlearned, change the underlying models' prediction. Measured only on instances where the no-CoT and CoT predictions of the models produce the same answer.

away from the initial prediction of the model (FF-SOFT). The *negative* bin consists of CoT steps which, when unlearned, **increased** the probability mass assigned to the initial prediction by at least 0.25. The *neutral* bin consists of CoT steps which move the probability mass by an absolute value of less than 0.25 in **either direction**. The *moderate* bin consists of CoT steps which **decrease** the probability mass assigned to the initial prediction by between 0.25 and 0.50. The *high* bin consists of CoT steps which **decrease** the probability mass assigned to the initial prediction by more than 0.50. We visualize the histogram of instances assigned to these bins in Figure 8.

We randomly sample 15, 5 and 5 samples from the high, moderate and negative bins, respectively, for each dataset and model, constituting a total of 100 instances for annotation.

Participants. We recruit a total of 15 volunteer participants to annotate the instances in the user

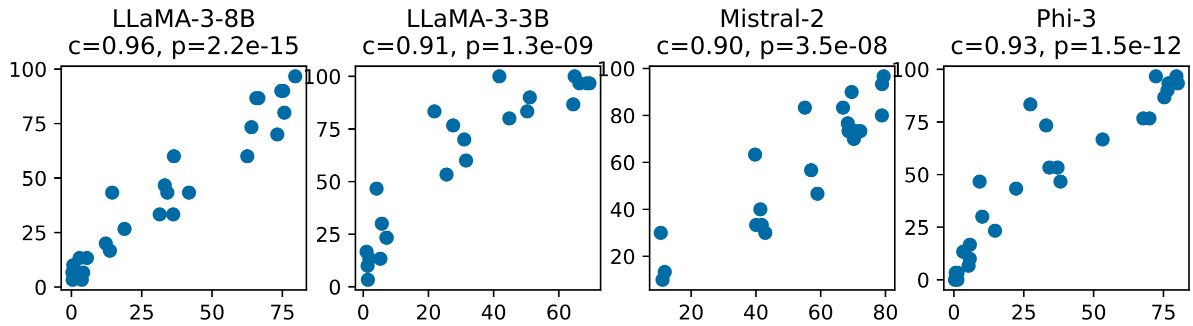

Figure 10: Scatter plot of correlation between efficacy and faithfulness, distributed across models. Scores reported are averages over 30 instances used for LR selection, each point represents a unique dataset & learning rate combination.

study, distribute the load equally between them and annotate each example once. All of the annotators are MA or PhD level students familiar with NLP. We use Qualtrics[11] to conduct the user study.

Protocol. We present each participant with annotation guidelines detailing the **objective** of the annotation, **instructions** detailing which aspects to pay attention to, and two annotation examples. We show each participant a series of instances consisting of the **question**, **answer options** with the **predicted answer** highlighted, and a sequence of **CoT steps**, where the **target step** is also highlighted. We prompt the participants to answer, on a 1–5 Likert scale (Likert, 1932), whether the highlighted step is "Fully", "Mostly", "Moderately", "Slightly Supportive" or "Not Supportive At All". We provide a screenshot from the annotation form in Figure 11.

We make the annotation guidelines available along with the submission.

K Hardware, Duration and Costs

Hardware Details We conduct our experiments on a computing system equipped with 32 Intel(R) Xeon(R) Gold 6430 CPUs operating at 1.0TB RAM. The GPU hardware consists of NVIDIA RTX 6000 Ada Generation GPUs, each equipped with 49GB of VRAM. Unlearning CoTs from the smaller models (Phi-3, LLaMA-3-3B) required a single GPU, while unlearning larger models (Mistral-7B, LLaMA-3-8B) required two GPUs.

Experiment Duration and Cost The initial implementation of unlearning experiments we conducted for an entire dataset took between 16 and 20 hours, depending on the model and dataset. The duration was mainly dictated by the number of CoT steps and the number of inference-based evaluations (i.e. generating CoTs post-unlearning, estimating specificity and CoT step probability for efficacy). The average duration of all full runs of models with final learning rates and exhaustive evaluation is 17h40m35s, with a standard deviation of approximately 1h56m38s.

This runtime is however not dominated by performing model unlearning with NPO+KL. When removing the various inference passes after each unlearning iteration which we used in the analysis, and just performing unlearning, the average runtime is 2h26m51s, with a standard deviation of 13m54s, representing a 7× speed-up, and highlighting that comprehensive evaluation used to report the full conducted analysis dominates the runtime.

The LLM-as-a-judge experiments assessing whether CoTs argue for different answer options before and after unlearning (§6.3) took between 6 and 8 minutes, per model and dataset. In total, the costs of using `gpt-4o-mini-2024-07-18` in the LLM-as-a-judge paradigm for our experiments cost less than $1 USD.

Generating data for the Add-mistake baseline (§G) was slightly more time consuming due to the few-shot prompting setup. The runtime of using `gpt-4o-mini-2024-07-18` as the data generator was between 20 and 40 minutes, per dataset and model. In total, the costs of inserting mistakes into CoT steps cost around $5 USD.

L Potential Risks

Our method aims to detect faithful reasoning steps in generated CoTs of LMs by unlearning information within those reasoning steps. We foresee two potential risks of our approach. Firstly, the faithful explanations detected by our model should not be

[11] https://www.qualtrics.com/

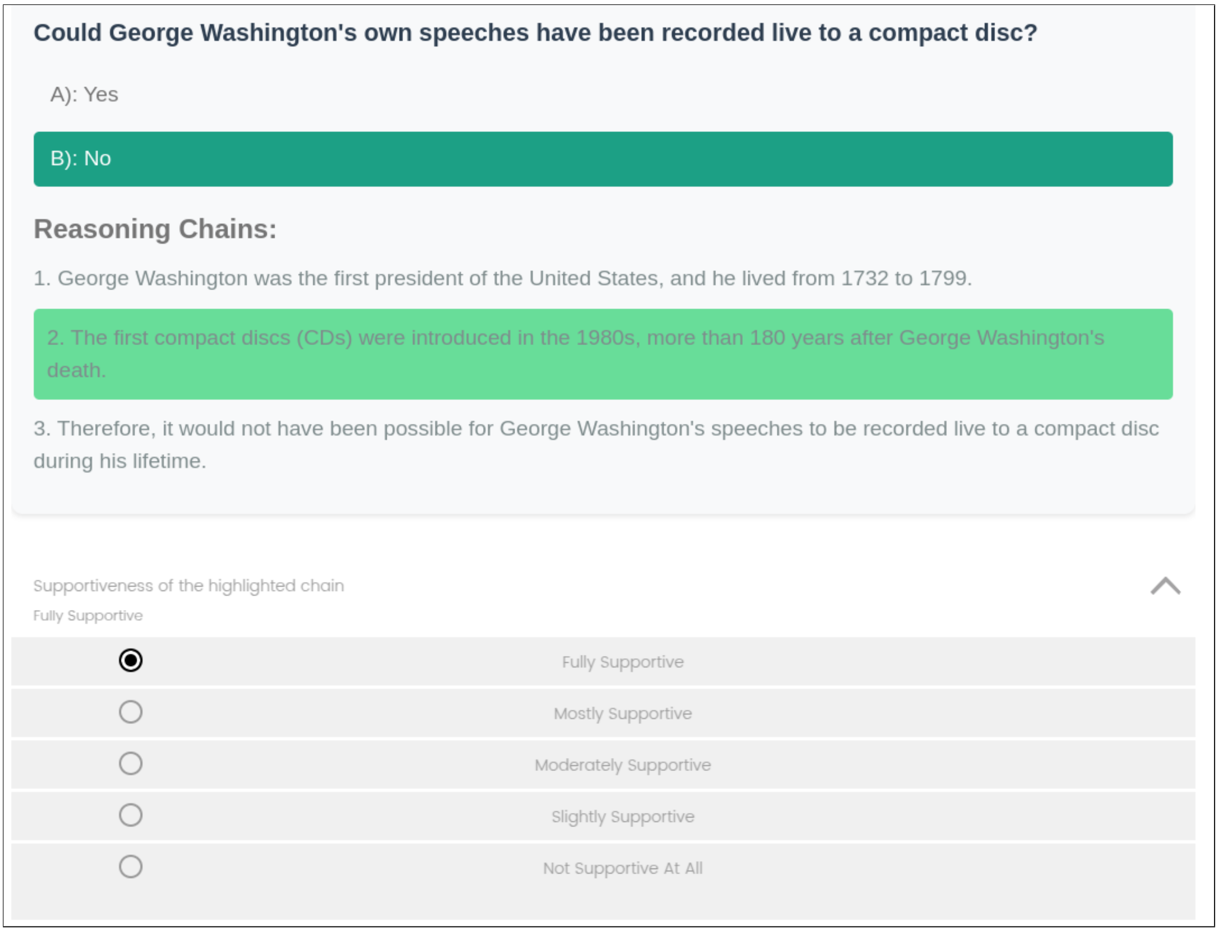

Figure 11: A screen capture of one example from the Qualtrics annotation platform. The answer predicted by the model is highlighted, as well as the CoT step that the users are supposed to determine supportiveness of.

taken as guidepoints for human reasoning. As our user study has shown (§6.4, §J), reasoning steps that are faithful to models are usually not plausible to humans, and should be used carefully in high-stakes scenarios. Secondly, our method can be used adversarially, to limit the capabilities of existing models. Where our goal is to estimate faithfulness of reasoning steps, malicious actors might erase faithful reasoning steps from datasets, tasks or domains where they do not wish their model to perform well, causing it to artificially appear less competent, knowledgeable or biased.