Investigating Mysteries of CoT-Augmented Distillation

Somin Wadhwa **Silvio Amir** **Byron C. Wallace**
Northeastern University
{wadhwa.s, s.amir, b.wallace}@northeastern.edu

Abstract

Eliciting "chain of thought" (CoT) rationales—sequences of token that convey a "reasoning" process—has been shown to consistently improve LLM performance on tasks like question answering. More recent efforts have shown that such rationales can also be used for *model distillation*: Including CoT sequences (elicited from a large "teacher" model) in addition to target labels when fine-tuning a small student model yields (often substantial) improvements. In this work we ask: **Why and how does this additional training signal help in model distillation?** We perform ablations to interrogate this, and report some potentially surprising results. Specifically: (1) Placing CoT sequences *after* labels (rather than before) realizes consistently better downstream performance—this means that no student "reasoning" is necessary at test time to realize gains. (2) When rationales are appended in this way, they need not be coherent reasoning sequences to yield improvements; performance increases are robust to permutations of CoT tokens, for example. In fact, (3) a small number of key tokens are sufficient to achieve improvements equivalent to those observed when full rationales are used in model distillation.

1 Introduction

Chain of thought (CoT) reasoning—i.e., generating tokens which communicate step-by-step "thinking"—can (sometimes dramatically) improve model performance on reasoning tasks (Wei et al., 2023). In the context of *model distillation* (Hinton et al., 2015), recent work has elicited such rationale chains from massive LLMs (e.g., GPT-4) to augment data with which to fine-tune much smaller (<2B parameters) task-specific models. Figure 1 illustrates this distillation approach: The student model is trained to generate the rationales in addition to the target token(s).

This simple CoT-augmented distillation strategy consistently and sometimes dramatically improves the performance of student models (Ho et al., 2023). For example, Li et al. (2023a) used rationales from GPT-3 (175B) to teach a comparatively tiny student LM (OPT-1.5B) to produce similar "reasoning" token sequences at inference time. They show an average increase in task accuracy of 12.4% across three commonsense reasoning datasets. Shridhar et al. (2023) adopted a similar approach to fine-tune GPT-2 (large; 774M) on grade-school math datasets with improvements of 8.23% on GSM8K and 16.20% on SVAMP. Beyond commonsense reasoning, Wadhwa et al. (2023) achieved SOTA results (+6.23 absolute gain in micro-F1, on average) with a distilled model for relation extraction by exploiting CoT rationales.

In this work we ask: ***Why* does distillation with CoT augmented targets consistently improve the performance of distilled LMs?** One might naively suspect that the student model benefits from learning to mimic the relevant "reasoning" process. But we find that it is not the case that student models benefit from "reasoning" at inference time.

Rather, consistent with contemporaneous work (Chen et al., 2024), we observe that placing CoT sequences *after* target tokens for distillation actually *improves* student performance (compared to when CoT is pre-prended to labels). This means the student model need not bother generating its "reasoning" at test time, as the label will be generated ahead of this anyway. Further, we find that rationale grammatically is not necessary; one can shuffle rationale tokens and/or include *only* "important" tokens from chains of thought during distillation and still realize performance benefits equivalent to those observed when using the full rationales.

Through ablations with three small student LMs (GPT-2, Phi-1.5, and Gemma-2B), we report the following, sometimes counter-intuitive, findings regarding CoT-augmented distillation and how rationales benefit student models. We summarize our key findings as

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6071–6086
November 12-16, 2024 ©2024 Association for Computational Linguistics

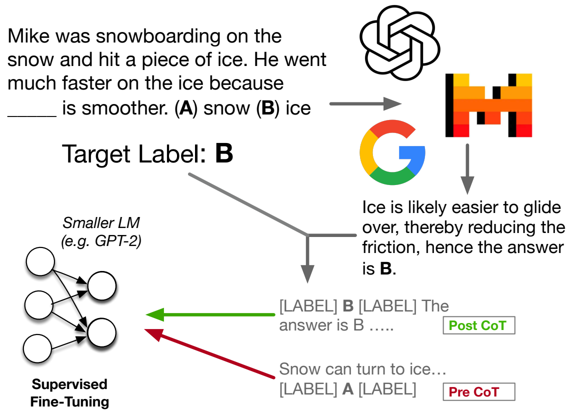

Mike was snowboarding on the snow and hit a piece of ice. He went much faster on the ice because _____ is smoother. (A) snow (B) ice
Target Label: B
Ice is likely easier to glide over, thereby reducing the friction, hence the answer is B.
Smaller LM (e.g. GPT-2)
[LABEL] B [LABEL] The answer is B .....
Post CoT
Snow can turn to ice...
[LABEL] A [LABEL]
Pre CoT
Supervised Fine-Tuning

Figure 1: For **RQ1**, we investigate augmenting CoT rationales obtained by very large (teacher) language models like Mistral, *after* the target labels. In doing so, we inject the same CoT reasoning ability during supervised fine-tuning (SFT) but do *not* condition generation of target label on the CoT itself at inference time.

1. **CoT-augmented distillation works better when rationales are provided *after* labels**. Standard CoT reasoning elicited zero-shot from massive LMs yields rationales as *prefixes* that logically lead to the label token(s). But we find that smaller models perform consistently *better* when rationales *follow* labels in distillation targets.

2. When *appended* to target labels, **token-level order, length, and coherence of rationales does not matter**. However, these things *do* matter when rationales are preprended. When the rationales are placed *before* the final label during fine-tuning, masking, shuffling, or altering coherent rationales significantly degrades model performance.

3. Motivated by the preceding observations, we run controlled experiments to establish that **there are certain key, contextual tokens that connect the input to the final label, and appending these tokens to labels is sufficient to achieve performance on-par with coherent CoT-like rationales**. It is solely the presence of these tokens at training time that leads to downstream performance improvements.

2 Experimental Design

CoT-augmented distillation entails eliciting rationales from a large *teacher* model and using these as additional training signal for a small *student* model. Rationales here comprise the logical steps taken to reach a response from a given input.[1] These are inserted into distillation training targets, and the student model is in this way taught to generate reasoning in addition to labels.

This has been shown empirically to provide (sometimes dramatic) performance benefits (Li et al., 2023a). But why? What accounts for the success of CoT-augmented distillation? In this work we investigate the following questions about the role of CoT-rationales in distillation. (**RQ1**) Does the placement of the reasoning chain relative to the target label (pre- or post-) matter? Relatedly, might observed performance gains owe to simply allowing the student model additional compute during inference? (**RQ2**) Must rationales feature logical and coherent "chain-of-thought" reasoning, or could we, e.g., scramble the ordering of tokens and still observe improvements? Finally, (**RQ3**) could we realize the same benefits in distillation using only a handful of key tokens from rationales, rather complete reasoning sequences?

To answer these questions empirically, we establish baseline student LM performance, and then compare this to ablated variants of CoT-augmented models. We use a fixed ICL prompt (Appendix B) with the input and target label to elicit a possible rationale for each instance in a dataset. We use Mistral-7B-Instruct (Jiang et al., 2023) as the teacher model and GPT-2 (Radford et al., 2019), Gemma-2B (Team et al., 2024) and Phi-1.5 (Li et al., 2023b) as student models. Note that one could instead replace Mistral-7B-Instruct with GPT-4 (or any other LLM capable of generating CoT-style rationales in ICL settings) as the teacher model. See Appendix B for the prompt used to elicit rationales for training instances of all datasets used in our work.

Following prior related work (Wei et al., 2023; Li et al., 2023a), we select three commonsense reasoning datasets: CommonsenseQA (Talmor et al., 2019), OpenbookQA (Mihaylov et al., 2018), and QuaRel (Tafjord et al., 2018). Each dataset provides an input consisting of a question, and a predefined set of answer choices. The target labels are the correct answer choices (Appendix A).

Implementation details We performed all of our experiments on two NVIDIA A100 GPUs. All student models (including ablations) were fine-tuned

[1] In the case of distillation, where one has access to reference labels, one can elicit rationales from the teacher which support the *correct* answer.

		CSQA	OBQA	QuaRel
Baseline (w/o CoT)	GPT-2	63.11	60.20	59.05
	Phi-1.5	67.77	56.81	76.82
	Gemma-2B	68.53	58.15	73.39
CoT before Label	GPT-2	67.20	69.71	66.27
	Phi-1.5	70.83	63.49	79.99
	Gemma-2B	70.61	65.85	74.90
CoT after Label	GPT-2	70.92	70.26	71.04
	Phi-1.5	72.56	72.49	81.36
	Gemma-2B	72.64	68.93	78.16

Table 1: Comparison of decoder-only models' performance under baseline supervised fine-tuning (no CoT), standard (pre) CoT, and postfix CoT.

with a learning rate of $3e$-5, batch size of 4 for CommonsenseQA and OpenBookQa, and 8 for QuaRel, with a maximum input length of 512, maximum output length of 256. We evaluated checkpoints every 500 steps with early stopping (patience $= 10$, threshold $= 0.02$). Because we are only interested in measuring relative performance of fine-tuned models across ablations (as opposed to necessarily realizing SOTA performance), we left the remaining hyperparamters to their default values.

RQ1: Positioning of Rationales

Does it matter if we place CoT rationales before or after target labels prior to distillation? Prior work (Wei et al., 2023) which elicited CoT reasoning from LLMs at inference time found that generating the chain *after* the final label performs comparably to the baseline (i.e., no CoT). This would seem to suggest that "reasoning" at inference time is what yields improvements, but it is unclear whether this holds in the context of model distillation. We compare the performance of student LMs distilled from examples with rationales placed both *before* and *after* labels (Figure 1).

`CoT before Label` `Friction is higher on rougher....[FIN_LABEL] B [FIN_LABEL]`
`CoT after Label` `[FIN_LABEL]` **`B`** `[FIN_LABEL] Friction is higher on rougher....`

We find that generating the CoT rationale *before* the label under-performs generating the rationale *after* the label. These findings are consistent across models and datasets (Table 1). Note that models trained to generate a CoT after the target label, do not need to do so at inference time. While in general CoT elicited from massive models is thought to improve performance by enabling explicit reasoning, gains offered in the context of distillation must be realized via some other mechanism (e.g., enriched training signal).

Next, we examine how and with what confidence do models fine-tuned under different conditions encode label information. We use ideas from LogitLens (nostalgebraist, 2020), TunedLens (Belrose et al., 2023), and FutureLens (Pal et al., 2023), which suggest that decoder-only models "think iteratively" and can be probed by inducing a distribution over the output vocabulary conditioned on hidden states to measure model confidence at different layers and time-steps within the model.

For each dataset, we look at test instances that are *correctly* predicted by all three model types, i.e., models distilled using: (i) No CoT; (ii) CoT before label; and (iii) CoT after label.

Figure 2 illustrates model confidences (i.e., probabilities computed with a softmax over the LM-head predictions for the final label) at different layers and time points, up to and including the final label prediction.[2]

In 80% of correctly predicted outputs, models trained with rationales *appended* to the final label (right-most subplot, Figure 2) correctly predict the label with probability > 0.6 at layer 32 and above. By contrast, models trained without any rationales (left-most subplot) lack such confidence, especially at lower layers: Final label probability does not exceed 0.6 until layer 44. Finally, for models trained with rationales *prepended* to target labels (middle sub-plot), the probability of the true label is ≤ 0.6 until layer 39 in 80% of correctly predicted instances. In sum, this analysis (illustrated in Figure 2) reveals a clear difference: Added CoT-information during distillation yields models which are more confident earlier on (positionally and layerwise) in the final output.

Is it *just* the extra "compute"? Prior work by Goyal et al. (2024) observed performance improvements in LLMs when inputs were augmented with "dummy" tokens (at pretraining and inference time), suggesting that LLMs benefit from additional compute cycles. Here we investigate whether it is just the added compute (i.e., steps/gradient updates over target label during fine-tuning) that provides gains *comparable to those achieved with CoT*, or if it is the CoT rationales contain useful information. Instead of CoT rationales, we prepend a fix-sized

[2] Full outputs omitted for brevity. See Appendix D for full length heatmaps.

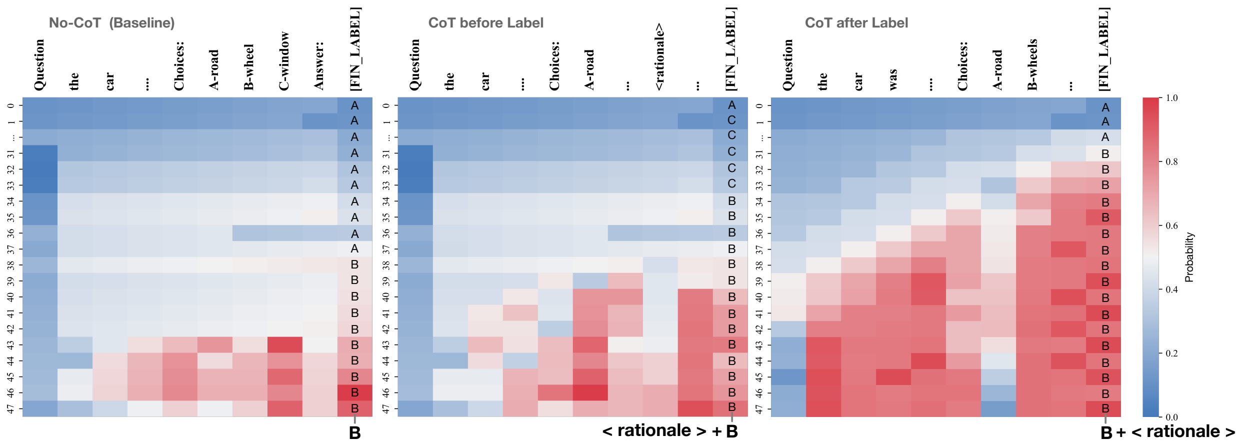


Figure 2: TunedLens (Belrose et al., 2023) visualizations on GPT-2 variants fine-tuned without CoT rationales (left), and with them pre-pended (middle) and appended (right). Augmenting distillation with CoT results in models that are more confident in labels earlier on. Models trained with rationales following labels are especially confident.

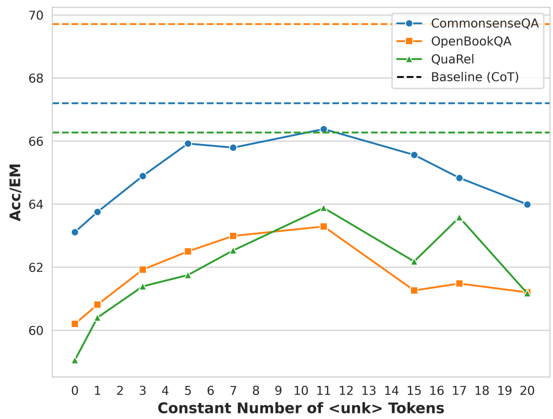


Figure 3: Performance of GPT-2 with constant number of <unk> tokens prepended to the target label.

sequence of <unk> tokens to the target label and ablate over the sequence length.

Figure 4 summarizes our results with GPT-2 as the student model; similar to Goyal et al. (2024) we observe that adding compute steps during training leads to (sometimes substantial) improvements in downstream performance. However, beyond a certain point ($\sim$ 11 <unk> tokens) performance plateaus, and then eventually declines. More importantly, at no point does the model outperform a CoT baseline (Table 1), suggesting that CoT rationales do indeed incorporate information necessary to achieve downstream improvements.

RQ2: Tokens within CoT Rationales

In light of our findings from RQ1, we next investigate what specific information in CoT rationales improves downstream performance. To this end, we assess how robust student models are to perturbations of provided rationales. Specifically, we consider: (i) Shuffling tokens within a rationale; and (ii) Incrementally masking tokens while retaining their relative order.

Shuffling We start by testing the robustness of student LMs with respect to the *coherence* of rationales. In particular, we shuffle tokens comprising rationales at the instance level. To illustrate this, consider the following example.

```
Question: If you hired a pitcher, (A) a
nerd (B) a bodybuilder, who likely can
pitch a baseball faster?
Original CoT Rationale: The answer is B
because bodybuilders typically have more
strength than nerds, which could
translate into a greater ability to
throw a baseball faster. [FIN_LABEL] B
[FIN_LABEL]
Shuffled Rationale: Baseball a faster
throw to ability greater a into
translate could which nerds than
strength more have typically
bodybuilders because B is answer The.
[FIN_LABEL] B [FIN_LABEL]
```

We then train the student LM with these shuffled rationales in place of the original (coherent) versions, under both pre- and post-label settings. Table 2 summarizes our findings from these experiments. We see that *prepending* the shuffled CoT rationales to target labels leads to sharp decline in performance, whereas *appending* these to target labels has nearly no effect on subsequent model

		CSQA	OBQA	QuaRel			CSQA	OBQA	QuaRel
CoT before Label	GPT-2	67.20	69.71	66.27	CoT after Label	GPT-2	70.92	70.26	71.04
	Phi-1.5	70.83	63.49	79.99		Phi-1.5	72.56	72.49	81.36
	Gemma-2B	70.61	65.85	74.90		Gemma-2B	72.64	68.93	78.16
Shuffled CoT before Label	GPT-2	34.56	41.64	32.88	Shuffled CoT after Label	GPT-2	69.56	70.15	70.56
	Phi-1.5	19.37	38.81	45.28		Phi-1.5	72.19	69.51	81.01
	Gemma-2B	25.80	35.17	20.52		Gemma-2B	71.13	67.28	76.50

Table 2: Comparison of model performance when shuffling the rationales to test for robustness to CoT coherence during SFT.

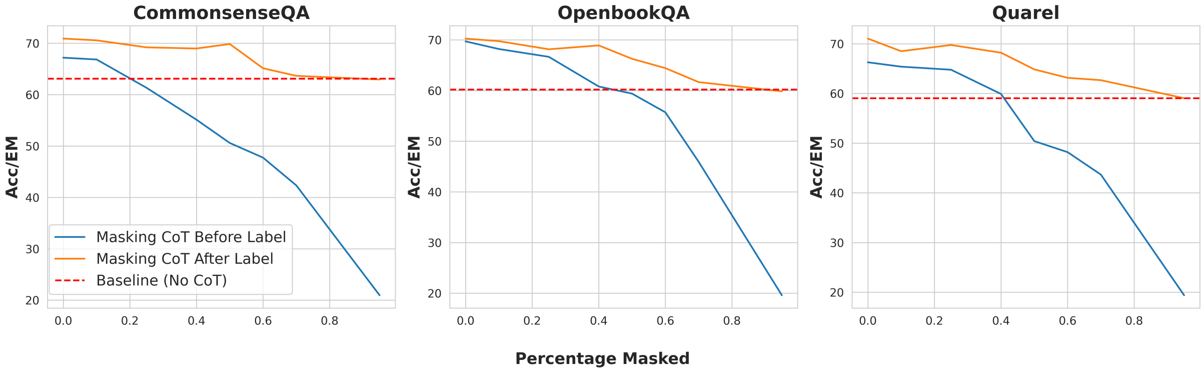


Figure 4: Comparison of model performance while successively reducing the amount of available information in a CoT rationale through masking.

performance.

Taken together with the results from RQ1, we hypothesize that this may be because prepending rationales to target labels during distillation requires the student model to learn to generate coherent rationales in addition to producing correct labels. By contrast, when rationales are appended they can serve as additional supervision during training without requiring coherent rationale generation at inference time.

Masking Next we run an ablation intended to test whether the full rationales are needed or if a subset of words is sufficient to realize the observed benefits. We start by randomly masking varying fractions of tokens within a rationale. For example:

Question: If you hired a pitcher, (A) a nerd (B) a bodybuilder, who likely can pitch a baseball faster?

Original CoT Rationale: The answer is B because bodybuilders typically have more strength than nerds, which could translate into a greater ability to throw a baseball faster. [FIN_LABEL] B [FIN_LABEL]

10% Masked: The answer is [MASK] because bodybuilders typically have more [MASK] than nerds, which could translate into a greater ability to throw a baseball [MASK]. [FIN_LABEL] B [FIN_LABEL]

50% Masked: The [MASK] is B [MASK] bodybuilders typically [MASK] more [MASK] than nerds, [MASK] could translate [MASK] a greater [MASK] to throw [MASK] baseball [MASK]. [FIN_LABEL] B [FIN_LABEL]

90% Masked: [MASK] [MASK] [MASK] B [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK], [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] faster. [FIN_LABEL] B [FIN_LABEL]

We vary the proportion of masked tokens from 10% to 90% (in increments of 10-15%) and again test under both pre- and post-label settings (see RQ1).

Figure 4 reports performances as a function of the proportion of masked tokens as compared to a non-CoT baseline. When only a small fraction (up to ∼20%) of rationale tokens are masked, we observe only marginal performance declines in both settings. However, as the proportion of masked CoT tokens increases (40%+), we see rapid performance decline in the CoT before label case —at

		CSQA	OBQA	QuaRel
Baseline (no CoT)	GPT-2	63.11	60.20	59.05
	Phi-1.5	67.77	56.81	76.82
	Gemma-2B	68.53	58.15	73.39
CoT after Labels	GPT-2	70.92	70.26	71.04
	Phi-1.5	72.56	72.49	81.36
	Gemma-2B	72.64	68.93	78.16
Grad Attr	GPT-2	71.30	74.86	71.26
	Phi-1.5	74.82	71.54	82.69
	Gemma-2B	73.85	68.13	79.03
Grad Attr Shuffled	GPT-2	71.24	74.99	71.47
	Phi-1.5	74.18	71.28	81.84
	Gemma-2B	72.93	67.30	78.94
Human Labels	GPT-2	–	–	67.06
	Phi-1.5	–	–	78.44
	Gemma-2B	–	–	74.77
Word2Vec Based	GPT-2	63.81	60.02	59.90
	Phi-1.5	67.94	56.22	75.49
	Gemma-2B	69.10	58.86	72.12

Table 3: Comparison of model performance under different attribution methods relative to retaining full length post-label CoT rationales.

60% masking, we find that the resultant distilled model performs worse than the baseline (i.e., without CoT).

We observe that masking a high percentage of tokens prior to the label yields models that generate a variable but often large number of [MASK] tokens prior to target label, often reaching maximum output length (set as a decoding hyperparameter). In contrast, in the CoT after label setting we observe gains over the non-CoT distillation baseline up until a high fraction ($> 60\%$) of tokens are masked; and beyond this point, the performance matches the vanilla (non-CoT) baseline.

RQ3: Attribution from Rationales

Having established that placing rationales after labels yields the best performance when performing CoT-augmented distillation—even without the full reasoning chain—we now ask whether we can find a small subset of "important" tokens that are sufficient to realize performance benefits. To determine importance, we consider both gradient-based attribution and human annotations.

Attribution via integrated gradients is a method to estimate the importance of individual tokens with respect to model output (Sundararajan et al., 2017). To measure the relative importance of rationale tokens on the final target label, we start with a baseline model (GPT-2) which is fine-tuned to generate a CoT-rationale *before* the final label. Considering a token sequence $\mathbf{x} = [x_1, x_2, \ldots, x_m, \ldots, x_n]$ where $x_{1,\ldots,m}$, correspond to the input tokens; $x_{m+1,\ldots,n-1}$ are the rationale tokens; and x_n is the final target label, we compute an approximation of the integrated gradients for the i-th rationale token as:

$$\mathrm{IG}_i \approx (x_i - x'_i) \sum_{k=1}^{p} \frac{\partial f(\mathbf{x}' + \frac{k}{m}(\mathbf{x} - \mathbf{x}'))}{\partial x_i} \cdot \frac{1}{p}$$

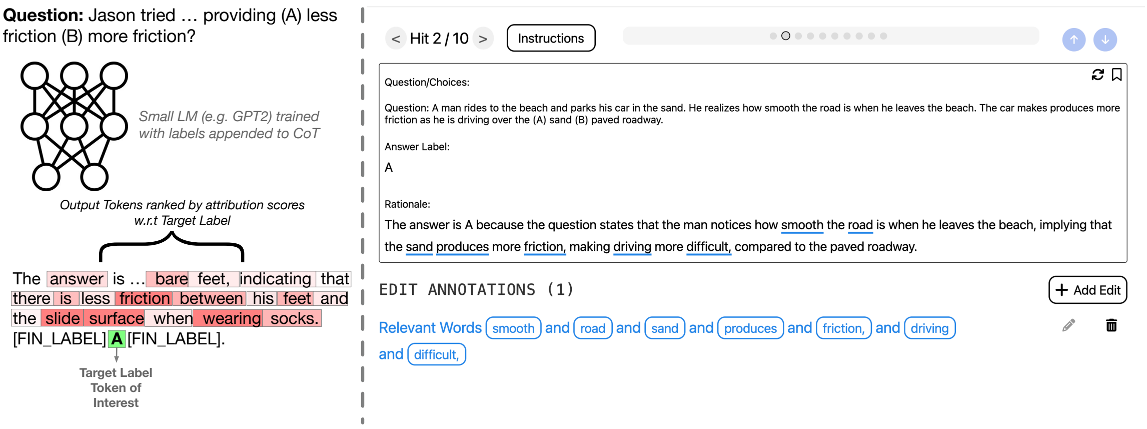
where, $\mathbf{x}'$ is a baseline (zero vector or a neutral input), $f(\mathbf{x})$ represents the model's output and p is the granularity for the approximation. See Sundararajan et al. 2017 for details.

The average length of a CoT rationale in our data is 36.3 tokens and we retain the top 15 tokens with the highest attribution scores. Figure 5 illustrates the process of computing a set of important tokens for the target label. As an example from QuaRel's training set[3]–

Question: If you hired a pitcher, (A) a nerd (B) a bodybuilder, who likely can pitch a baseball faster?
Original CoT Rationale: [FIN_LABEL] B [FIN_LABEL] The answer is B because bodybuilders typically have more strength than nerds, which could translate into a greater ability to throw a baseball faster.
Attributed Tokens: [FIN_LABEL] B [FIN_LABEL] B because body builders more strength translate throw baseball faster
Shuffled Attributed Tokens:[FIN_LABEL] B [FIN_LABEL] translate because more body B faster builders baseball strength throw

We fine-tune the models again, *appending* the attributed tokens to the final label, similar to the case where CoT rationales are generated after the label at inference. Table 3 summarizes our findings. We broadly observe that reducing the total number of tokens in CoT rationale via gradient attribution leads to no significant difference in downstream model performance. This is consistent with our findings in RQ2, where masking a majority of CoT tokens *appended* to the target label did not significantly effect model performance.

[3] Training instance ID: V1_B3_0128


Question: Jason tried … providing (A) less friction (B) more friction?
Small LM (e.g. GPT2) trained with labels appended to CoT
Output Tokens ranked by attribution scores w.r.t Target Label
The answer is … bare feet, indicating that there is less friction between his feet and the slide surface when wearing socks. [FIN_LABEL] A [FIN_LABEL].
Target Label Token of Interest
< Hit 2 / 10 >
Instructions
Add Edit

Question/Choices:

Question: A man rides to the beach and parks his car in the sand. He realizes how smooth the road is when he leaves the beach. The car makes produces more friction as he is driving over the (A) sand (B) paved roadway.

Answer Label:

A

Rationale:

The answer is A because the question states that the man notices how smooth the road is when he leaves the beach, implying that the sand produces more friction, making driving more difficult, compared to the paved roadway.

EDIT ANNOTATIONS (1)

Relevant Words smooth and road and sand and produces and friction, and driving and difficult,

Figure 5: Comparison of Attribution Methods: Left side we have automated extraction via Integrated Gradients while the right side displays manually annotated words *perceived* by human annotators to be relevant.

Human annotations As an alternative to scoring rationale tokens via gradient attribution, we evaluate using tokens that humans *perceive* to be the most relevant to target labels. We hire annotators on Prolific[4] to identify minimal sets of (up to 15) words in rationales necessary to answer the question (Figure 5).[5] We first ran a small internal pilot to estimate time required and set fair pay rates.[6] Next we collected annotations for ∼2k instances of the QuaRel dataset in batches of 200 from crowdworkers fluent in English. We manually verified 10% (20 instances) of each batch to ensure quality.[7]

Replacing CoT rationales with words deemed important by annotators offers some gains, but smaller than those from integrated gradients (Table 3). To measure overlap between tokens selected manually and via gradient attribution, we assume the latter to be the reference tokens and measure Precision (0.73) and Recall (0.59) of annotated words[8] In general, we find the set of tokens identified through gradient attribution to be much more comprehensive than those selected by human annotators.

Are tokens "similar to" the label sufficient? Finally, we explore whether CoT rationales elicited from LLMs merely provide tokens that are similar to but distinct from target labels. One way to collect such tokens is to select from a large set of words a subset that have high similarity to the target label token. To this end we use static Word2Vec (Mikolov et al., 2013) embeddings.[9]

We select the 15 closest words to the target label for all training instances. For target labels with multiple tokens, we take similarity with respect to only the longest token. We then use these retrieved tokens in lieu of CoT rationales when fine-tuning student models. The question is whether the additional information encoded in words similar to the target label yield performance gains comparable to CoT augmented distillation. Table 3 reports results, which are largely negative for this experiment: The observed performance with "relevant word augmentation" is comparable to the no-CoT setting (baseline). This suggests that while rationales need not be coherent to realize benefits, the tokens they comprise must offer additional signal beyond being simply "similar" to (in terms of co-occurence) target label tokens.

3 Related Work

Distillation via elicitation West et al. (2022) considered "symbolic" distillation where instead of distilling from soft representations like logits, they proposed the use of LLMs as data generators to be used to augment training data. Other recent work has shown that *explanations* can serve as both inputs (Hase and Bansal, 2022) and targets (Wiegreffe et al., 2022), and can be used downstream to improve task- (Wadhwa et al., 2023) and domain-specific (Ho et al., 2023) model performance.

[4] `https://www.prolific.com/`

[5] Interface designed using `https://thresh.tools`.

[6] We pay US$15/hr to all crowdworkers regardless of their geographic location.

[7] We required all crowdworkers to have an overall job approval rating of ≥95% with at least 100 completed jobs on Prolific.

[8] We assume a complete overlap if any subword in an annotated word matches with a reference token.

[9] `word2vec-google-news-300`[10]; trained on the Google News dataset of ∼100 billion words.

Li et al. (2023a) first explored distillation performance to tasks like commonsense reasoning and provided analyses intended to reveal factors that may be important in creating the *teacher* corpus, upon which our work builds on. Beyond directly using explanations-style rationales for fine-tuning, Deng et al. (2023) explored an alternative approach by using model hidden states to perform *implicit* reasoning, instead of producing rationale tokens one-by-one (i.e. Next Token Prediction), demonstrating that the chains of thought themselves may not be fully necessary to achieve downstream fine-tuning performance gains.

Our work deepens these efforts by focusing on analyzing specific fine-tuning for distillation dynamics in smaller models, and characterizing when rationales generated by teacher LLMs are helpful.

CoT with Small Models In-context CoT prompting (Wei et al., 2023) induces thinking *step-by-step*, such that the model generates intermediate reasoning ultimately leading to a target label. Prior work (Ho et al., 2023; Magister et al., 2023) has shown that small models may not be not inherently capable of generating these reasoning chains, but can be taught to do so using augmented training sets.

Creating CoT-augmented training sets can be expensive, and a number of prior works in the area have investigated synthetic data generation. For instance, Hsieh et al. (2023) generate new target labels from few instances of labeled data. Li et al. (2023a) notably found that sampling multiple rationales can improve small-model performance. Han et al. (2023) decomposed the reasoning steps into multi-round dialog and optimize for the correct path using PPO algorithm while training smaller models. Fu et al. (2023) emphasize the trade-offs between task-specific CoT-generation capability in small models and their generalizability. Wang et al. (2023) establish the effect of faithfulness of elicited rationales on the student models trained using them. A shared theme in these past papers have been that they explicitly look at improving the *quality* of rationales themselves and its downstream effects on overall model performance.

Our work differs from these efforts looking only at the final label, manipulating the CoT rationales *at distillation (fine-tuning) time* to probe how rationales effect model performance.

Contemporaneous Work While engaged in this work, a few contemporaneous efforts have surfaced which make some observations that overlap with our findings. Chen et al. (2024) introduce "post-semantic thinking" (PST) to reduce the influence of rationales on final output labels. Xu et al. (2024) reveal that preemptive answer generation (a target label) within a CoT rationale is highly sensitive to malicious attacks, which comports with our hypothesis (i.e. vice versa) that a faulty reasoning leading to incorrect rationales can effect the overall model performance (which is solely evaluated on labels generated after those rationales).

4 Conclusions

We have investigated *why* and *under what circumstances* does CoT-augmented distillation improve student model performance. Specifically, we evaluated the degree to which the following aspects contribute to the observed gains realized in CoT-augmented distillation.

1. The placement of rationales (before or after labels). **Finding: Appending (rather than prepending) rationales to targets yields consistently better performance.**

2. The coherence of rationales and their grammatically. **Finding: When rationales follow labels, the words they comprise can be scrambled and one still observes comparable gains.**

3. Whether we need only a small set of key tokens from rationales (and how to identify them). **Finding: Gains comparable to CoT-augmented distillation can be realized using a small set of tokens identified via gradient attribution; using manually selected "important" words does not do as well, nor does using tokens that are "similar to" label words.**

Some of these findings corroborate and deepen observations made in contemporaneous work, e.g., models can benefit from additional compute at inference time (Goyal et al., 2024), and CoT-augmentation fares best when rationales are placed *after* the target labels (Chen et al., 2024). We have not fully characterized the mechanism by which CoT augmentation aids distillation, but we have ruled out some explanations and provided empirical insights into when and how CoT augmentation provides useful signal to student models.

Limitations

There are important limitations to this work and the conclusions we can draw from it.

First, we have only considered publicly available open-domain question-answering datasets in our analyses, to the exclusion of complex information extraction tasks such as relation extraction where CoT-augmented distillation has also proven useful (Wadhwa et al., 2023). We made this choice largely in the interest of consistency with prior work, and to avoid complex evaluation challenges that occur during generative relation extraction.

Second, we did not attempt to *improve* the quality of CoT rationales generated by teacher models through iterative prompt refinement or other techniques (Wang et al., 2023). We also elicited the rationales for distillation from modestly sized open source models, rather than (for example) GPT-4. It may be possible to elicit "better" rationales from massive proprietary models, but it seems unlikely (though possible) that our conclusions vis-a-vis distillation would change as a result.

Third, our evaluation of using rationale tokens annotated manually is limited by the way we framed the task. It could be that an alternative design and/or annotation interface would yield different annotations, and this may in turn lead to different conclusions regarding the utility of tokens selected in this way.

Finally, we *only* experimented with English-language datasets and we therefore cannot say whether these results would hold in other languages.

Ethics Statement

This work required some human annotations which were collected using an online platform called Prolific. Prolific required us to pay workers *per hour*, and so we had to estimate the time required to complete one batch of annotations. To do so, we (the authors) carried out a small number of these annotations to determine the approximate hourly compensation. We then set the compensation rate to average $15 USD/hour. If annotators took longer than expected to complete a batch of annotations, we paid bonuses to ensure that their cumulative pay averaged out to US$15/hour.

Statement of Intended Use Our work relies on open source datasets and models. Like any trained model, there is a risk of the distilled model inheriting or amplifying any biases present in the original LLM's rationales. While rationales make the model more interpretable than a blackbox classifier, there still may be challenges in fully explaining the distilled model's behavior. While distilling, the user must be aware of these considerations and institute appropriate safeguards.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) grant IIS-1901117.

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens.

Xiaoshu Chen, Sihang Zhou, Ke Liang, and Xinwang Liu. 2024. Post-semantic-thinking: A robust strategy to distill reasoning capacity from large language models.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2023. Implicit chain of thought reasoning via knowledge distillation.

Yao Fu, Hao-Chun Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *ArXiv*, abs/2301.12726.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens.

Chengcheng Han, Xiaowei Du, Che Zhang, Yixin Lian, Xiang Li, Ming Gao, and Baoyuan Wang. 2023. Dialcot meets ppo: Decomposing and exploring reasoning paths in smaller language models.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39, Dublin, Ireland. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander J. Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *ArXiv*, abs/2305.02301.

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.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023a. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks are all you need ii: phi-1.5 technical report.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

nostalgebraist. 2020. interpreting gpt: the logit lens.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328. JMLR.org.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen tau Yih, and Ashish Sabharwal. 2018. Quarel: A dataset and models for answering questions about qualitative relationships. In *AAAI Conference on Artificial Intelligence*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–

15589, Toronto, Canada. Association for Computational Linguistics.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rongwu Xu, Zehan Qi, and Wei Xu. 2024. Preemptive answer "attacks" on chain-of-thought reasoning.

Appendix

A Dataset Details

We conducted our experiments using three datasets; for completeness we provide details about these here.

CommonsenseQA (Talmor et al., 2019) is a multiple-choice question answering dataset that requires commonsense knowledge. Each question is accompanied by five answer choices; only one is correct. The dataset consists of 12,102 questions split into a training, development, and test sets of set of 9,741, 1,221, and 1,140 questions, respectively. The following is an example from the training data (ID: `7e93dacd4d1b7c7aa4c15f5da220bd59`)

Question: The two conglomerates decided to reach tentative agreement to what?
Choices:
A: do business
B: accomplish
C: stop arguing
D: make progress
E: digging holes
Answer: A (do business)

OpenBookQA (Mihaylov et al., 2018) is designed to test an understanding of elementary science, combining factual knowledge with commonsense reasoning. The dataset contains 5,957 questions, each with four answer choices (and one correct response). This is split into training, development, and test sets of 4,957, 500, and 500 questions respectively. A unique aspect of OpenbookQA is its focus on scientific facts which students are expected to know. The following is an example from the training data (ID: `12-271`)

Question: Skills are learned characteristics. To get better at doing something, you must stretch yourself in ways that
Choices:
A: may be very uncomfortable at first
B: take very little time
C: are without learning from others and past experiences
D: are without goals and commitment
Answer: A (may be very uncomfortable at first)

QuaRel (Tafjord et al., 2018) is a dataset for reasoning over physical processes involving comparative relationships. It consists of 2,740 multiple-choice questions, each with two answer choices (one being correct). The questions require reasoning about how physical processes affect different entities in qualitative ways. The dataset provides train/development/test splits comprising 1,948/278/514 questions. The following is an example from the training data (ID: `QuaRel_V1_Fr_0344`)

Question: Ryan races his car and needs to drive in different types of situations. Ryan drives around in a sandy desert, and then in an empty parking lot. After each drive, Ryan sees how warm his car got. Ryan notices that his car was much warmer after driving in the sand than it was after driving in the parking lot. That is because the sand had _____ than the parking lot.
Choices:
A: more resistance
B: less resistance
Answer: A (more resistance)

B Prompts

Our experiments required eliciting chain-of-thought (CoT) rationales from a "teacher" LLM to be used in distillation. For this we used Mistral-7B-Instruct (Jiang et al., 2023). We used the following rationale-augmented few-shot prompt to this end. The question, answer choices, and the target label are taken from the original training instance, and the CoT rationale provided was written by us (the authors).

CommonsenseQA

<s>[INST] Given the following two examples of question-answer-rationale triplets, provide a rationale for the third example for why the selected choice answers the question. [\INST]
Question: The president had to make a decision regarding the hate attack on his country, what did he do? **Choices:**A: wage war; B: fight enemy; C: kill; D: destroy enemy; E: attacked his country
Answer: A (wage war)
Rationale: The answer is A because the president's decision to address a hate attack on his country typically involves taking military action, such as waging war, to protect and defend the nation. </s>
Question: Letters are sometimes delivered by hand through one of these? **Choices:**A: mail box; B: suitcase; C: front door; D: bowl; E: post office
Answer: C (front door)

Rationale: The answer is C because letters are delivered by hand through the front door.</s>

OpenBookQA

<s>[INST] Given the following two examples of question-answer-rationale triplets, provide a rationale for the third example for why the selected choice answers the question. [\INST]
Question: Oak tree seeds are planted and a sidewalk is paved right next to that spot, until eventually, the tree is tall and the roots must extend past the sidewalk, which means **Choices:**A: roots may fall apart; B: roots may begin to die; C: parts may break the concrete; D: roots may be split;
Answer: C (parts may break the concrete)
Rationale: The answer is C because as the oak tree grows, its roots may exert pressure on the sidewalk, causing the concrete to crack or break.</s>
Question: A cow eats some hay, an apple and a piece of bread. In its tummy **Choices:**A: mail box; B: suitcase; C: front door; D: bowl; E: post office
Answer: B (suitcase)
Rationale: The answer is B because the cow's stomach contains digestive enzymes that break down the consumed food into smaller, soluble molecules through the process of dissolution.</s>

QuaRel

<s>[INST] Given the following two examples of question-answer-rationale triplets, provide a rationale for the third example for why the selected choice answers the question. [\INST]
Question: Dan drives a car into the garage from the gravel parking lot. The car moves more smoothly into the garage than the parking lot. This is because there is a bumpier surface in the (A) garage floor (B) gravel parking lot.
Answer: B (gravel parking lot)
Rationale: The answer is B because the question states that the car moves more smoothly into the garage than the parking lot, indicating that the gravel parking lot has a bumpier surface compared to the garage floor. </s>
Question: he baseball team coach was considering both Ted and Roy for the right field position. He needed someone who could propel the ball all the way to the basemen and he knew Ted was more physically fit and possessed greater physical strength than Roy. Who could likely throw the ball a further distance? (A) Roy (B) Ted
Answer: B (Ted)
Rationale: The answer is B because the question indicates that Ted is more physically fit and possesses greater physical strength than Roy, suggesting that Ted is more likely to throw the ball a further distance.</s>

C Models and Reproducibility

We used the Huggingface library (v4.26.1; Wolf et al. 2020) and publicly available checkpoints for both student[11] and teacher[12] models. GPT-2 and Phi-1.5 were fine-tuned on a single A100 instance while Gemma-2B was fine-tuned on 2 A100 instances. To monitor the training process, we evaluated model checkpoints every 500 steps. Early stopping was employed with a patience parameter of 10, meaning that training was halted if there was no improvement in the evaluation-set accuracy for 10 consecutive evaluations. The improvement threshold was set to 0.02, ensuring that only significant improvements were considered to continue training. This strategy helped to prevent overfitting and reduced unnecessary computational overhead. Upon publication, we release all code (included elicited rationales for all datasets considered) necessary for reproducing our experiments.

D RQ1 example heatmaps

Now we visualize the predictions of individual layers of GPT-2 fine-tuned with no, pre, and post-CoT rationales while processing the input "Question: an electric car contains a motor that runs on...Choices: A: gas; B: hydrogen; C: ions; D: plutonium", we are specifically interested in what the layers in the penultimate time-step (w.r.t final target label) *think* the next token should be.

[11] https://huggingface.co/openai-community/gpt2-xl; https://huggingface.co/microsoft/phi-1_5;https://huggingface.co/google/gemma-2b

[12] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

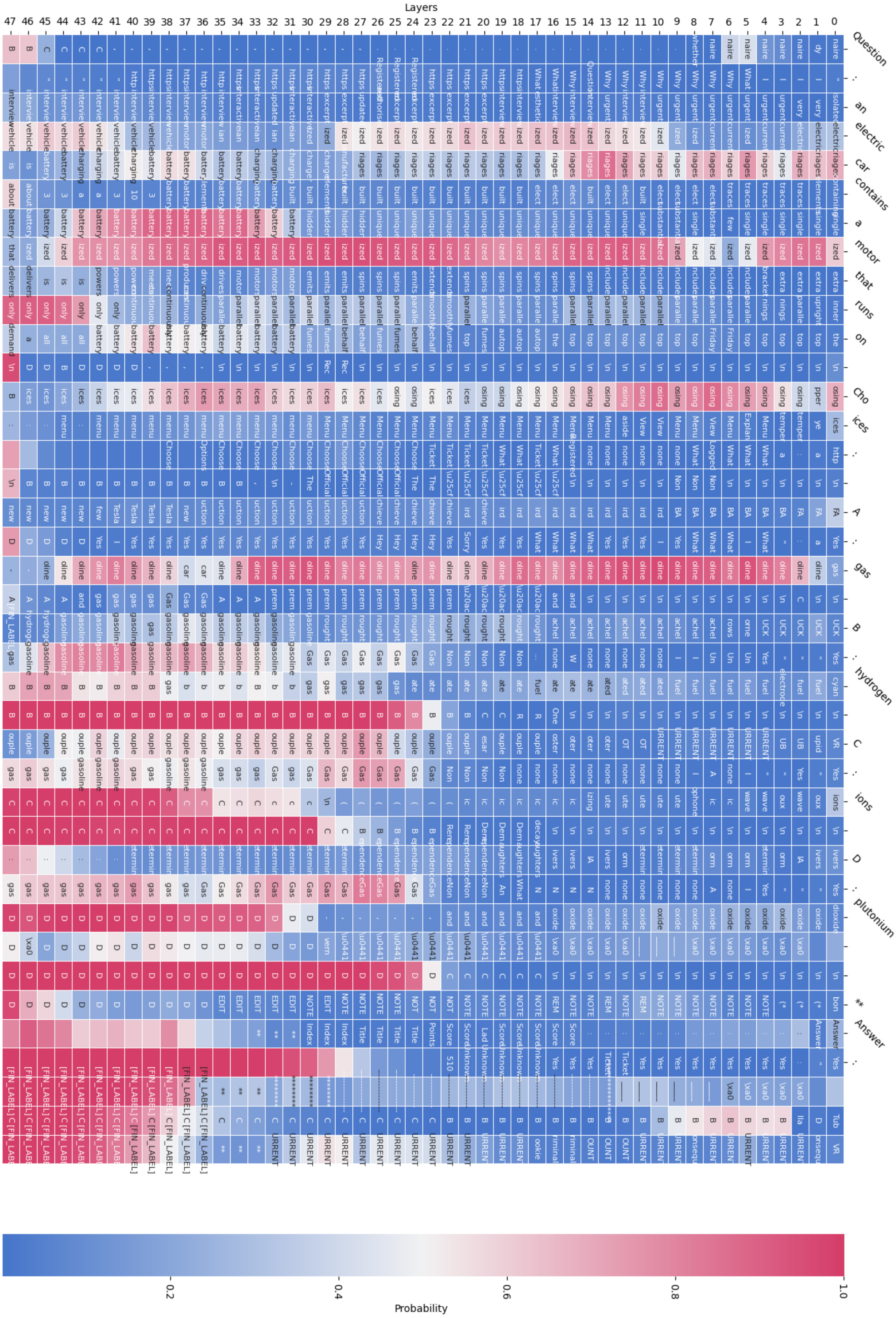


Figure 6: **No-CoT Baseline**: GPT-2 variant fine-tuned without COT rationales. In this example, the model is confident (p>0.6) of the first occurrence of the correct label C at **layer 40**.

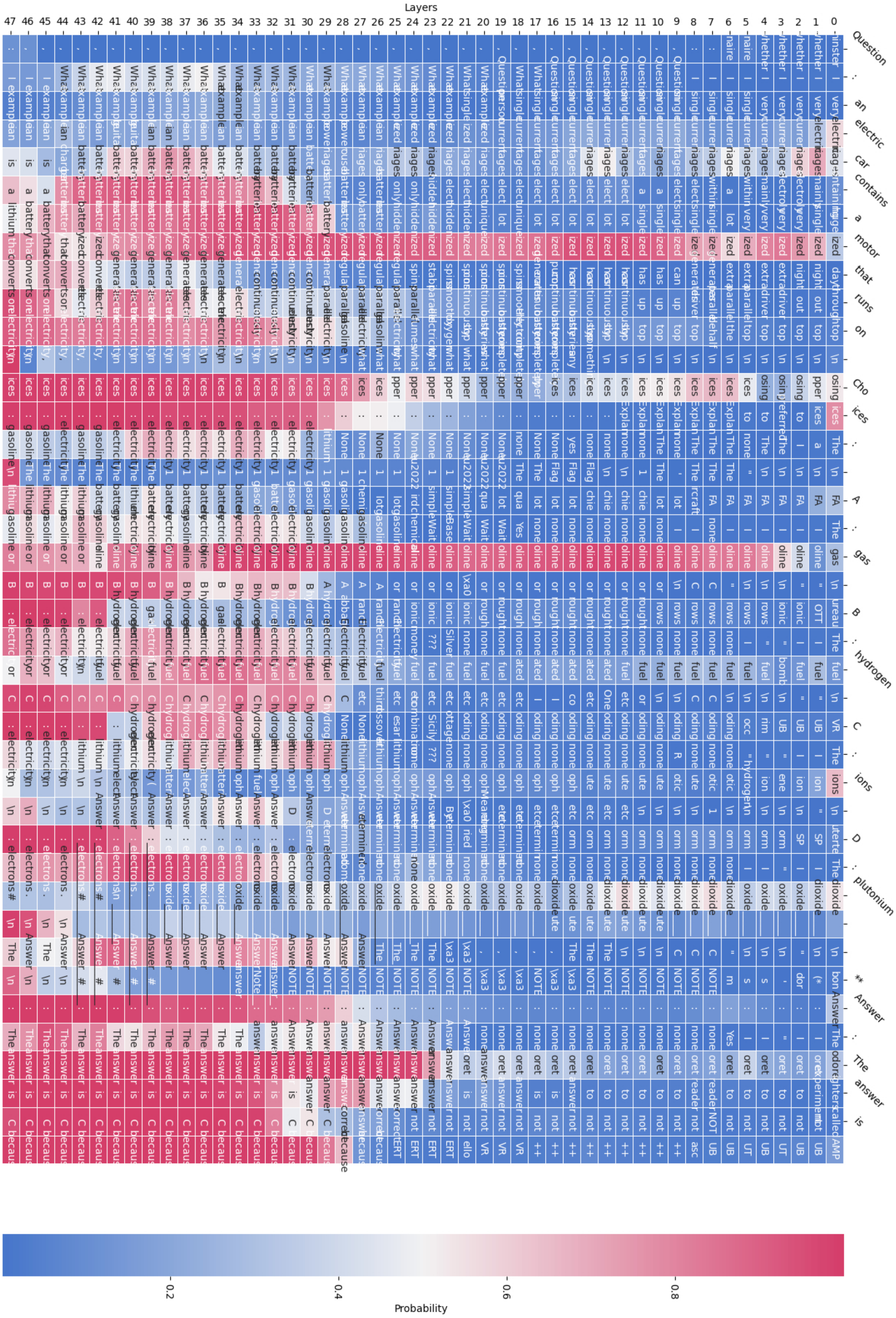


Figure 7: **Pre-CoT**: GPT-2 variant fine-tuned with CoT rationales *pre-pended* to the target label. In this example, the model is confident (p>0.6) of the first occurrence of the correct label C at **layer 33**.

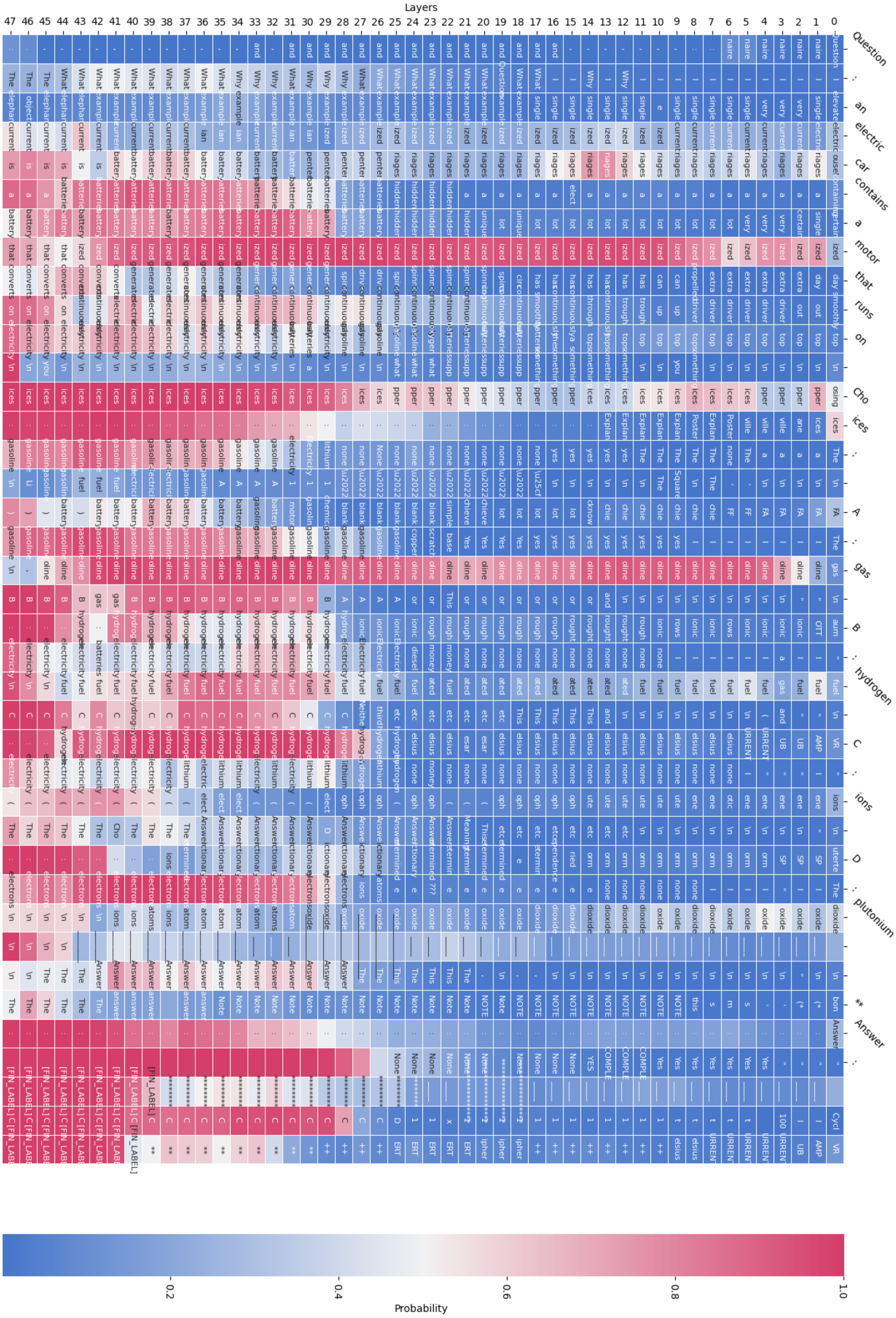


Figure 8: **Post-CoT**: GPT-2 variant fine-tuned with CoT rationales *appended* to the target label. In this example, the model is confident (p>0.6) of the first occurrence of the correct label C at **layer 27**.