Rationales Are Not Silver Bullets: Measuring the Impact of Rationales on Model Performance and Reliability

Chiwei Zhu[1], **Benfeng Xu**[1,2§], **An Yang**[2], **Junyang Lin**[2]
Quan Wang[3], **Chang Zhou**[2†], **Zhendong Mao**[1]
[1]University of Science and Technology of China
[2]Qwen Team, Alibaba Inc
[3]Beijing University of Posts and Telecommunications
{tanz, benfeng}@mail.ustc.edu.cn

Abstract

Training language models with rationales augmentation has been shown to be beneficial in many existing works. In this paper, we identify that such a prevailing view does not hold consistently. We conduct comprehensive investigations to thoroughly inspect the impact of rationales on model performance as well as a novel perspective of model reliability. The results lead to several key findings that add new insights upon existing understandings: 1) Rationales can, at times, deteriorate model performance; 2) Rationales can, at times, improve model reliability, even outperforming their untrained counterparts; 3) A linear correspondence exists in between the performance and reliability improvements, while both are driven by the intrinsic difficulty of the task. These findings provide informative regulations on the broad utilization of rationales and raise critical implications on the procedure of explicitly aligning language models with implicit human thoughts. Codes can be found at https://github.com/Ignoramus0817/rationales.

1 Introduction

It is widely acknowledged that the capabilities of large language models can be significantly enhanced when they are given time to *think*, a process where a **rationale** is generated first to mimic human inner thought before arriving at the final answer (Wei et al., 2023; Kojima et al., 2023). Although this concept was initially identified and established in the context of prompting large language models (LLMs), it has also been extensively explored in training language models (LMs) as well. In general, rationales have profoundly influenced our understanding and utilization of LMs.

Benefiting from the powerful capabilities of LLMs (OpenAI et al., 2023), it is now much more approachable to obtain high-quality, large-scale synthetic reasoning traces as rationales. As a result, **Rationale-Augmented finetuning (RAFT)** has been receiving increasing attention in many recent works, where a rationale is concatenated between the original question and answer to augment the learning process of a weaker or smaller language model (Figure 1). RAFT has brought consistent benefits for diversified tasks including mathematical reasoning (Shridhar et al., 2023; Magister et al., 2023), question answering (Wang et al., 2023a; Li et al., 2022), symbolic reasoning (Magister et al., 2023) as well as general-purpose chatbots (Li et al., 2023; Mitra et al., 2023). As a result, adding rationales is becoming a default measure when finetuning smaller models recently.

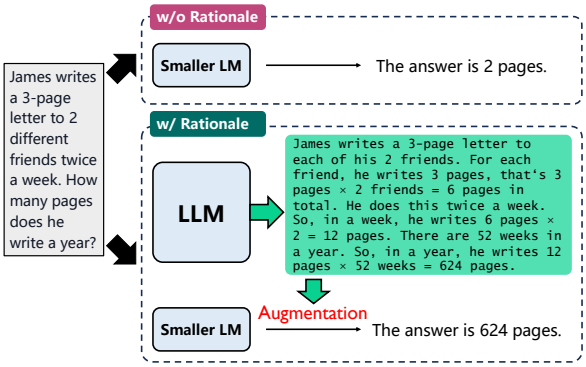

Figure 1: Illustration of training LMs with rationale augmentation.

In this paper, we identify some dissonance in the broad beneficial effect of RAFT, where introducing rationales can have a negative impact on model performance on certain tasks. Through far more comprehensive and extensive investigations, we present results that diverge from previous expectations. Aside from model performance, we include **Calibration** as an essential supplemental perspective of our analysis. Calibration refers to whether a model can assign its predictions with proper probabilities that reflect the actual likelihood of its results

§Corresponding author. Work done during the internship at Alibaba Group.
†Work done while working at Alibaba Group.

Findings of the Association for Computational Linguistics: ACL 2025, pages 5808–5835
July 27 - August 1, 2025

being correct (Guo et al., 2017). It indicates the reliability of a deep model and can be affected when the prediction process is augmented with rationales.

Collectively, we measure the impact of RAFT on model performance and reliability on a total of 18 tasks. Figure 2 provides an overview of all results, and key findings can be summarized as follows.

Key Findings I

On Performance: Rationales indeed bring variable benefits for many difficult tasks, but this does not invariably hold across all tasks (corresponding to area $x < 0$).

In 11 out of 18 tasks, performance has dropped when augmented with rationales, which is beyond expectation: one would expect RAFT to at least do no harm if it does not bring much improvements. It might be noticed that recent `gpt-4-o1` model employs rationales for better performance, seemingly challenging this work. Actually, it helps strengthen our conclusions. On the one hand, capabilities of `gpt-4-o1` mainly exhibits in math and code reasoning fields, which is consistent with our observations. On the other hand, it demonstrates that more sophisticated effort should be made to make better use of rationales.

Key Findings II

On Reliability: Model calibration error can benefit from rationales, and even outperforms its untrained base model in certain tasks ($y > 0$).

Studies have pointed out that pretrained language models are well calibrated enough and finetuning process would degrade the calibration of language models (Guo et al., 2017; Kadavath et al., 2022; He et al., 2023; Zhu et al., 2023). However, when such a finetuning process is augmented with rationales, this degradation can be alleviated (12 out of 18 tasks). There are 3 tasks where calibration error under RAFT is even slightly better than its untrained base model.

Key Findings III

Performance-Reliability Correlation: There exists a significant linear correspondence between the improvement in model performance and reliability under rationale-augmented training ($y = \alpha \cdot x + \beta$).

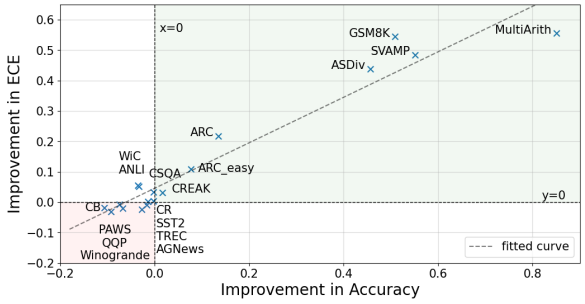

Figure 2: Improvement in Accuracy (x-axis) and Expected Calibration Error (ECE) (y-axis) under RAFT for different datasets.

We empirically find that rationale's impacts on performance and reliability are synchronized. We attribute this linear correlation to the intrinsic difficulty of specific tasks. We further propose several difficulty metrics to validate this assumption and establish, for the first time, a quantitative relationship between correlation and task difficulty.

We also design extensive ablations to verify the robustness of our findings across varied conditions. Finally, we locate the reason of the impacts of rationales and provide explanations with qualitative study. In appendices we also present exploratory analysis and discussions, delving deeper into the underlying mechanism of rationales and their impacts. In general, this paper depicts a systematic view of the impacts of rationale-augmented finetuning, revealing a deeper understanding as well as new insights into its utilization and mechanisms.

2 Preliminaries

2.1 Rationale-Augmented finetuning

Rationales. Rationales are free-text reasoning steps produced by either human beings or language models when solving problems. Specifically, for language models, rationales can be defined as reasoning steps produced ahead of answers in Chain-of-Thought inference (see Fig. 1).

Rationale-Augmented Finetuning. RAFT is introducing rationales in the finetuning process. We conduct RAFT in the following steps: Given a supervised dataset $D = \{x_i, y_i\}_{i=1}^{n}$ where x_i and y_i are the i_{th} input and label, rationales r_i are first generated for each sample, resulting in the augmented dataset $D' = \{x_i, (r_i, y_i)\}$. Then we finetune a language model on the augmented dataset maximizing the probability of generating $(r_i; y_i)$, where ";" means concatenating. For comparison, we train the

model using the same hyper-parameters but with solely answer labels, i.e. maximizing the probability of generating y_i. Then we compare the performance and calibration of the two models. To measure model performance and calibration quantitatively, we use **Accuracy (Acc)** and **Expected Calibration Error (ECE)** as metrics following previous works (Guo et al., 2017; Li et al., 2023).

2.2 Calibration

Confidence Calibration. Given a classification problem where we have the input $X \in \mathcal{X}$, label $Y \in \mathcal{Y} = \{1, 2...K\}$, which are random variables following ground truth joint distribution. Also we have model prediction $Y' \in \mathcal{Y}$ and assigned confidence $P' \in [0, 1]$. A model is perfectly calibrated if it satisfies the following equation:

$$P(Y' = Y|P' = p) = p, \quad \forall p \in [0, 1] \quad (1)$$

However, the probability in Equation 1 is not calculable. To measure model calibration statistical approximations are often used, including Expected Calibration Error (Naeini et al., 2015).

Expected Calibration Error. Expected Calibration Error (ECE) is a quantitative measurement of calibration using finite data samples. Given a set of N predictions and their confidence, we divide the confidence interval $[0, 1]$ into M bins with equal length $1/M$ and group these predictions according to their confidence. Then we can calculate the average accuracy and confidence of each bin:

$$Acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}_i = y_i), \quad (2)$$

$$Conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \quad (3)$$

where B_m is the set of indices of samples in the m_{th} bin. $\hat{y}_i$ and y_i are prediction and ground truth for the i_{th} sample. $\mathbb{1}$ is an indicator function generating 1 if the prediction is correct and 0 otherwise. $\hat{p}_i$ is the confidence that the model assigns to the i_{th} prediction. Then ECE is calculated as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |Acc(B_m) - Conf(B_m)| \quad (4)$$

ECE measures the gap between average confidence and accuracy among all bins. Lower ECE means better calibration. We use $M = 10$ in all experiments, following previous works (Guo et al., 2017; Desai and Durrett, 2020; He et al., 2023).

2.3 Datasets and Implement Details

Datasets We conduct our research over 18 datasets in 7 categories, including Math Reasoning, Common Sense Reasoning, Sentiment and Topic Analysis, Paraphrasing, Natural Language Inference, Word Sense Disambiguation and Coreference Resolution. Dataset details for each category can be found in Appendix B.1.

Rationale Generation For each dataset, we construct a training set of 20k samples augmented with rationales. As obtaining large-scale human annotations is costly, we utilize `gpt-3.5-turbo-0613` to generate rationales following (Fu et al., 2023). For each data point, we formulate the input with manually written prompts and query `gpt-3.5-turbo-0613` to generate rationales, from which we only keep rationales that lead to correct answers*. We recursively traverse the dataset until we get enough data. Prompts for rationale generation are in Appendix D. Though keeping only rationales leading to correct answers has already guaranteed the data quality to some extent, we further conduct a brief quality examination to make sure the generated rationales, which confirm the quality of the data(details in Appendix C).

Training and Inference. We selected LLaMA-2-7B (Touvron et al., 2023) as our base model. Full hyper-parameters for training are in Appendix E. During inference, we apply a self-consistency voting (Wang et al., 2023b). For every input question, we sample 10 reasoning paths with a temperature of 0.8, each generating an answer, from which the one that appears most frequently is kept as the final answer. If the answer appears n times in the sampled results, we consider the confidence as $n/10$.

3 Impacts of Rationales on Performance and Reliability

We record the accuracy and ECE difference of models finetuned with and without rationales as follows:

$$\Delta Acc = Acc_{RAFT} - Acc_{FT}, \quad (5)$$

$$\Delta ECE = -(ECE_{RAFT} - ECE_{FT}), \quad (6)$$

Note that we use a negative increase of ECE as lower ECE means better calibration.

*For each sample we repeatedly generate until getting the correct answer or reaching the retry limit 10.

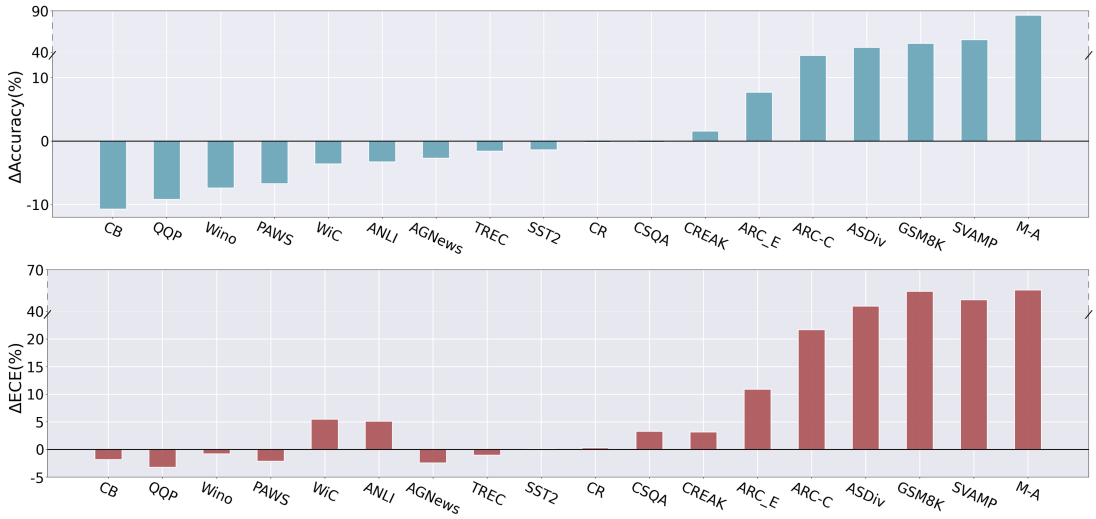

Figure 3: Improvements in accuracy and ECE under RAFT. Datasets are re-ordered according to the improvements in accuracy. Wino refers to Winogrande and M-A refers to MultiArith. Note that y-axes are folded for better display.

Settings		GSM8K	MultiArith	ASDiv	SVAMP	ARC-Challenge	Average
Untrained (ICL)	① w/o Rationale	0.180	0.215	0.077	0.082	0.057	0.122
	② w/ Rationale	0.082	0.251	0.054	0.039	0.100	0.105
	② - ①	-0.098	+0.036	-0.023	-0.043	+0.043	-0.017
Finetuned	③ w/o Rationale	0.581	0.612	0.501	0.525	0.313	0.506
	④ w/ Rationale	0.037	0.057	0.063	0.041	0.096	0.059
	④ - ③	-0.544	-0.555	-0.438	-0.484	-0.224	-0.447
Δ FT-Base	③ - ①	+0.401	+0.397	+0.424	+0.443	+0.256	+0.384
	④ - ②	-0.045	-0.194	+0.009	+0.002	-0.004	-0.046

Figure 4: ECE of finetuned and base models. $\Delta_{FT-Base}$ means difference of finetuned and pretrained models.

3.1 Impacts on Model Performance

As is seen in Fig. 3, the most significant improvement in accuracy happens in the math reasoning task, where accuracies of all four datasets improve by more than 40%. Rationales also raise model performance on the two ARC datasets. Surprisingly, rationales do not always bring performance gain. Instead, they distract models from getting the correct answers for most of the tasks.

3.2 Impacts on Model Reliability

As can be seen in Fig. 3, in most tasks models under RAFT are better calibrated than their counterparts finetuned with answer labels. Fig. 4 shows the results of tasks where model calibration improves the most. Firstly, from the table we can see that finetuning with labels does much harm to model calibration (③ - ①). Then it is noticeable that for both untrained and finetuned models, incorporating rationales brings benefit to model calibration (② - ①, ④ - ③). Lastly, in 3 tasks, models with RAFT can achieve even better ECE than the untrained models (④ - ②). Our results verify the established conclusion that finetuning will damage model calibration, while also showing that introducing rationales can alleviate such harmful effects.

3.3 Linear Correlation Between Impacts on Performance and Reliability

As can be seen in Fig. 2, the improvement in accuracy and ECE across different datasets show a significant linear correlation, depicted as follows:

$$\Delta ECE = \alpha \Delta Acc + \beta, \tag{7}$$

where $\alpha = 0.7479, \beta = 0.0456$. Line fitted with Ordinary Least Squares algorithm (Galton, 1886) can be seen in Fig. 2. The Pearson Coefficient and p-value from the significance test are 0.9681 and $2.462e^{-10}$ respectively, portraying a very distinct linear correlation. Such correspondence may also explain why models become worse-calibrated for tasks where accuracy drops the most.

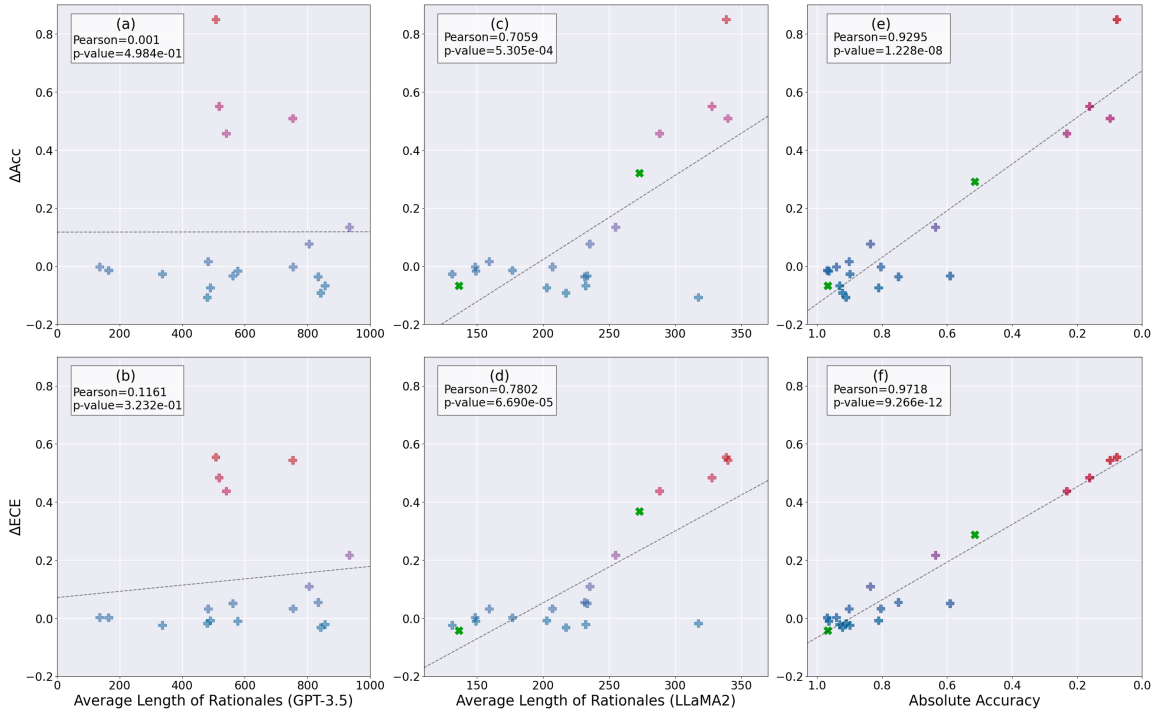

Figure 5: Correspondence between model improvement and task difficulty across different metrics: (a), (b): Average length of rationales from `gpt-3.5-turbo-0613`. (c), (d): Average length of rationales from LLaMA-2-7B-base. (e), (f): Absolute Accuracy of LLaMA-2 model finetuned with answer labels (x-ticks is reversed). The color of the points shifts from blue to red as Y-values increase.

4 Analyze Impacts of Rationales with Difficulty

It is notable that improvements are more significant for intuitively harder tasks, e.g. ARC-Challenge compared to ARC-Easy. Thus we assume the impacts of rationales on performance and calibration are both related to the inherent difficulty of tasks, which can explain the linear correspondence. However, correctly defining the intrinsic difficulty of a task is non-trivial. In this section, we explore three alternative metrics to approach a reasonable characterization of task difficulty as follows.

4.1 Definition of Task Difficulty

Metric 1: Lengths of Rationales Generated by GPT-3.5. It is well-acknowledged that the duration of human thought is positively correlated with problem difficulty (Kotovsky et al., 1985; Kahneman, 2011). We propose that a similar relationship exists for pretrained language models, as they possess extensive general world knowledge. This parallel between rationale length and computational effort is frequently drawn in contemporary research on inference-time scaling (Snell et al., 2024). As a further justification, we demonstrate on a subset that it shows distinct linear correlation with an established dataset difficulty metric, V-usable Information (Ethayarajh et al., 2025), with a Pearson coefficient of 0.895 (Details in Appendix F. Note that this method requires training a separate model for each dataset to measure, so we only use it for a preliminary study to demonstrate the effectiveness of our metrics).

So we use the average length of rationales generated by `gpt-3.5-turbo-0613` as a measurement for task difficulty. Specifically, for each test set, we sample 100 data points, apply a zero-shot Chain-of-Thought inference to each of them using `gpt-3.5-turbo-0613`, and collect the rationales. For each piece of data, we repeatedly generate 10 times, resulting in 1,000 rationales for the dataset, which we call **reference rationales** in the following parts. Then average length is calculated to measure the difficulty of this dataset.

Metric 2: Lengths of Rationales Generated by LLaMA2-base. As GPT-3.5 is trained with closed-source SFT data, the generated rationales might potentially suffer from certain biases. For instance, its SFT data may contain math questions similar to those in GSM8k, thus impacting the length of generated rationale as well as its neutrality as a difficulty metric. In order to pur-

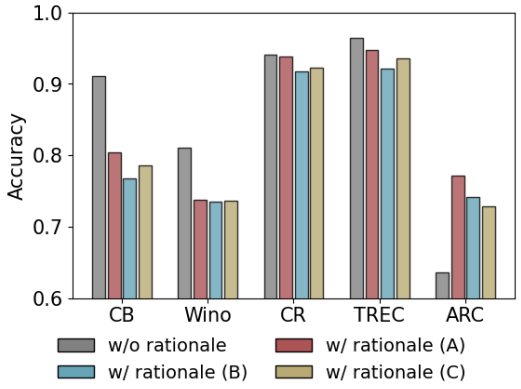

Figure 6: Impacts of varied rationales.

sue unbiased rationales that naturally arises from pre-trained LLMs, we design a **Mix-of-Task In-Context Prompting** strategy. We first employ LLaMA2 as the generator, since it is not explicitly aligned to human instructions, we prepend 3 demonstrations to regulate its behavior to generate rationales for each given instance. The crucial aspect here is to make sure all three demonstration tasks are different from the target task so that the generated rationale is neither biased toward the alignment procedure nor biased towards any of the demonstrations. The prompts are constructed as:

$$P = \{(x_i, r_i, y_i)^{t_1}; (x_j, r_j, y_j)^{t_2}; (x_k, r_k, y_k)^{t_3}; x^{t_n}\} \quad (8)$$

where $(x_i, r_i, y_i)^{t_1}$ are input, rationale and answer sampled from task t_1. x^{t_n} is the input from target task t_n. Demonstration tasks t_1, t_2, and t_3 are randomly sampled from the task pool excluding t_n.

Metric 3: Absolute Accuracy. Another alternative metric is the absolute accuracy of models. We assume that lower accuracy corresponds to higher task difficulty as models are less capable of these tasks. We consider the accuracy under vanilla label-only finetuning for consistency.

4.2 Attributing Impacts of Rationales with Difficulty

Fig. 5 shows how the improvement in accuracy and ECE evolves with the increase in task difficulty. Clear rising trends are evident in Fig. 5 (c), (d), (e), and (f), in which reference rationales generated by LLaMA-2 and absolute accuracy are used to measure difficulty. Meanwhile, we find the correlation between accuracy improvement and task difficulty bears a resemblance to that between ECE and task difficulty. Such a trend may indicate that the improvements in accuracy and ECE are similarly driven by task difficulty, thus resulting in a linear correlation with each other. Note that when using GPT-3.5 generated rationales as references, there is no significant positive relation, which might be caused by the bias we mentioned before.

4.3 Actionable Insight: Estimating Gain of RAFT with Task Difficulty

The above findings in this paper directly points out an actionable insight: we can estimate the effects of RAFT based on various difficulty-related factors. Although strictly calculate the specific impacts is challenging, we manage to measure the difficulty of tasks with lengths of LLaMA2 generated rationales and absolute accuracy of models, which provides a preliminary solution and may serve as actionable guidance. To be specific, the linear correspondence in Fig. 5 can be formulated as follows:

$$\Delta Acc = 0.0029 \times Len(R_{LLaMA2}) - 0.5567 \quad (9)$$

$$\Delta Acc = 0.8031 \times Acc_{FT} + 0.6730 \quad (10)$$

Based on the equations above, it is predictable with meaningful confidence whether and to what extent the performance and reliability improves when RAFT is applied for a given task. We verify the above relationship with two other datasets, SUBJ and CoinFlip (details in Appendix B.1), which are green crosses in Figure 5(c-f). We believe our findings serve as a good initiation in the field of generating and using rationales more predictably. Additionally, though this work mainly focuses on task-specific training, our conclusions may generalize to a broader topic, the alignment of large language models, which we discuss in Appendix A.2.

5 Robustness of the Impacts Brought by Rationales

As model performance and calibration can be affected by several factors, we conduct ablations to verify the robustness of our conclusions.

Model selection. Extra experiments are conducted on Qwen-7B, LLaMA2-7B-Chat, and LLaMA2-13B. Additionally, we adopt another two stronger models as rationale generators, LLaMA-3.1-70B-Instruct and GPT-4o-mini (results in Appendix G). We perform RAFT on 6 datasets where performance varies in our previous experiments. As is shown in Fig. 8 and Fig. 9, rationales still

harm certain tasks across all models. And again, significant linear correspondence can be identified between the improvement in ECE and accuracy, despite that the slopes and intercepts are different.

Hyper-parameters. We conduct a search on learning rate (from $5e^{-7}$ to $5e^{-4}$) and training epochs (from 2 to 6 epochs). Accuracy and ECE from the best models are reported (results in Appendix G). We select 3 datasets which are representative of the typical effect of RAFT: harm, improvement, or unchanging. In Table 6 we can see that it remains unchanged whether RAFT brings improvement or harm to model performance and calibration on all datasets. Consequently, our conclusions about the impact of RAFT are robust against different models and hyper-parameters.

Different Prompts. Rationales are generated with prompts from another two annotators (see Appendix D.3) and are used for RAFT. Results are displayed in Fig. 6. For all tasks we investigate, newly trained models behave the same way as the original ones, where RAFT uniformly improves or damages models' performance, which demonstrates the robustness of our key finding I. Key finding II is similarly verified consistent across different annotators in Appendix H.

Multi-task Training. We conduct multi-task training to verify our findings in a out-of-distribution scene. We utilize two methods of constructing training data for multi-task training:

- Polarity Mixture: Separately mix data from datasets where RAFT brings gain (or harm).
- Full Mixture: Mix data from all datasets.

We train models with mixed data and report test performance on QQP, CR, and GSM8K. Details and results can be found in Appendix K. As shown in Fig. 15, none of the multi-task models behaves differently from singlet-task models. As a result, multi-task instruction tuning does not change whether RAFT would cause improvement or harm.

Rationale-Augmented Prompting. We investigate whether these conclusions still hold under the frequently used inference paradigm, Rationale-Augmented Prompting. We query `gpt-3.5-turbo-0613` to generate answers for the test sets with and without rationales. Still not all tasks benefit from rationales. However, the previous findings remain valid: incorporating rationales continues to improve ECE, with accuracy and ECE improvements showing a linear relationship. Finally, although there is a positive correlation between improvements and task difficulty, it is less pronounced than that observed in the RAFT settings (see Table 10), which may indicate that the effect of rationales is less influenced by task difficulty in the RAP setting. Full results are in Appendix J.

6 Explanatory Analysis

In this section we try to explain the impacts observed above. We first exclude the formality of RAFT and the extra computation as potential causes through a blank rationale experiment and then conduct a qualitative analysis to identify situations where rationales cause performance drop. Furthermore, despite the difficulty of fully explaining our conclusions from the perspective of inner mechanisms of RAFT, we provide some hypotheses and a discussion in the Appendix A.1.

6.1 Locating the Cause of Impacts

Given that rationales are additionally inserted between the question and answer, an intuitive argument would be that these reasoning traces should at least not deteriorate model performance if they do not provide obvious benefits. We thus design an ablation experiment where rationales are replaced with equal-length blank sequence $\{\texttt{<Think\_i>}\}^L$ (See Appendix I for sequence example). This ablation can effectively disentangle the meaning and extra computation brought by rationales.

Results are shown in Appendix I, where Fig. 11 shows that blank rationales indeed neutralize the negative impacts of rationales, bringing performance in line with label-only finetuning as expected. We can thus exclude either the formality of RAFT or the extra computation as potential causes, and the real cause for performance deterioration should be the inherent meaning of the rationales.

6.2 Qualitative Analysis

To further investigate the performance drop, we collect 100 samples where models under RAFT generate wrong answers while models trained with labels do well. We conclude 3 types of errors as shown in Table 1. **`Incorrect Rationales`** refers to the rationales that directly lead to wrong answers. Specifically, **`Unreasonable`** means the generated rationales are logically wrong. As shown by the example, the rationale is opposite to the

Table 1: Typical cases of different errors. Blue marks correct reasoning traces or golden answers and red marks oppositely.

Error Attribution		Datasets	Samples
Incorrect Reasoning (75%)	**Unreasonable** (57%)	Winogrande	**Question:** At night, Jeffrey always stays up later than Hunter to watch TV because _ wakes up late. **Answer:** ...Jeffrey stays up later to watch TV because the other person wakes up late. Therefore, the most likely option is Hunter. (Correct: Jeffrey)
	Off-Topic (18%)	AGNews	**News:** Sun Microsystems Inc. has decided to postpone its September SunNetwork 2004 San Francisco user conference, and ... **Answer:** ...The overarching topic of the news is related to business, specifically the decisions and contemplations made by Sun Microsystems Inc. regarding their conferences. The news is about Business. (Correct: Science/Technology)
Trivial Reasoning (23%)		TREC	**Text:** What position did Willie Davis play in baseball ? **Answer:** This question is about the position played by Willie Davis in baseball. So it is about an Entity. (Correct: Human Beings)
Incorrect Conclusions (2%)		Winogrande	**Question:** Adam was law-abiding and not a crook like Joseph, so it was surprising that _ was the one with a restraining order against him. **Answer:** It implies that the person who was law-abiding and not a crook like Joseph is the one who had a restraining order against them. So the answer is Joseph. (Correct: Adam)

meaning of the question. In such cases, models only mimic the form of reasoning steps but are not indeed reasoning. `Off-topic` is a different situation, where the rationales make sense, but miss the key point to solve the problem. In the example, the rationale focuses on certain words, ignoring the topic itself. These cases are naturally difficult for models, where introducing rationales may not produce better results. `Trivial Rationales` are those rationales that just repeat the problem. This usually happens for tasks with simple forms like sentiment/topic classification, where conclusions can be jumped to without much reasoning content. Finally, `Incorrect Conclusions` are rare cases where generated answers are not consistent with the reasoning steps, which might be occasional and we do not further look into.

7 Related Work

Rationales are free-text reasoning steps produced by human or language models. Much attention has been paid to leveraging rationales to enhance model performance (Zaidan et al., 2007; Druck et al., 2009; Zhang et al., 2016; Camburu et al., 2018). Recently LLMs have been capable of generating reasoning steps through Chain-of-Thought (Wei et al., 2023; Kojima et al., 2023), which inspires a bunch of works studying rationale-augmented finetuning (RAFT) using LLMs (Nye et al., 2021; Chung et al., 2022; Fu et al., 2023; Shridhar et al., 2023; Mukherjee et al., 2023; Mitra et al., 2023). These works focus more on proposing new methods of extracting and utilizing rationales on certain target tasks. There are also works studying the effectiveness of rationales (Carton et al., 2022; Hase and Bansal, 2022; Yao et al., 2023; Kabra et al., 2023). Yet none of the above has studied the effect of rationales widely on different tasks nor do they study the effect of rationales on calibration (While (Sprague et al., 2024) introduces similar idea to ours, they mainly focuses on zero-shot setting, and do not study the behavior of calibration). Our study works in this untraveled direction.

Confidence calibration is first proposed for determining how a weather forecaster is reliable (Miller, 1962; Murphy, 1973). Research on confidence calibration of statistical machine learning methods has a long history (DeGroot and Fienberg, 1983; Palmer et al., 2008), and later calibration of neural networks is also researched on (Nguyen and O'Connor, 2015; Hendrycks and Gimpel, 2016; Nixon et al., 2019). (Guo et al., 2017) points out that finetuned neural networks are over-confident, which is potentially caused by over-minimizing the loss. Recently, calibration of language models also received attention. (Desai and Durrett, 2020; Kadavath et al., 2022) study the calibration of pretrained language models and

point out that they are well calibrated. Similarly, finetuning has been found to harm the calibration of LMs (He et al., 2023; OpenAI et al., 2023; Zhu et al., 2023). While previous works have studied the calibration of pretrained and finetuned language models, we expand the calibration measurement to the rationale-augmented finetuning setting.

8 Conclusion

In this work, we systematically examine the impact of rationale-augmented finetuning (RAFT) on model performance and reliability, and bring out several key findings that add new insights to current understandings. It sometimes deteriorates model performance, while alleviating calibration error caused by finetuning over-fitting. We also identify a significant linear correlation between the impacts on performance and reliability, both are driven by the intrinsic difficulty of the task. With exploratory analysis and discussions, this paper implies future directions to continually delve into the underlying mechanism of rationales, pursuing better alignment between the explicit reasoning process of auto-regressive language models and implicit human thought structure.

Acknowledgment

This research is supported by Artificial Intelligence-National Science and Technology Major Project 2023ZD0121200 and National Natural Science Foundation of China under Grant 62222212.

Limitation

Our work reveals several new findings about rationale-augmented finetuning through extensive experiments. Although this work presents explanations and discussions about the intrinsic mechanism of rationales, theoretical proof is still pending. We hope this work would inspire new attempts on more rigorous formulation for rationale and its inherent mechanism on language models.

Potential Risks

Our work focuses on the mechanism and effects of rationales, which provide guidance on the prevailing solution of rationale-augmented finetuning, and might also inspire improved methods to produce better synthetic rationales when training language models. Negative impact may include the abuse of LLMs to generate rationales for malicious tasks or using improved language models for harmful content generation.

References

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Preprint*, arXiv:1812.01193.

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to learn, and how: Toward effective learning from rationales. *Preprint*, arXiv:2112.00071.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

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. *ArXiv*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. *Preprint*, arXiv:2003.07892.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240.

DrewWham and Mauricio Nascimento. 2020. Coin flips. Kaggle.

Gregory Druck, Burr Settles, and Andrew McCallum. 2009. Active learning by labeling features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 81–90, Singapore. Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2025. Understanding dataset difficulty with $\mathcal{V}$-usable information. *Preprint*, arXiv:2110.08420.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *Preprint*, arXiv:2301.12726.

Francis Galton. 1886. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *Preprint*, arXiv:1706.04599.

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.

Guande He, Jianfei Chen, and Jun Zhu. 2023. Preserving pre-trained features helps calibrate fine-tuned language models. *Preprint*, arXiv:2305.19249.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Anubha Kabra, Sanketh Rangreji, Yash Mathur, Aman Madaan, Emmy Liu, and Graham Neubig. 2023. Program-aided reasoners (better) know what they know. *Preprint*, arXiv:2311.09553.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

K Kotovsky, J.R Hayes, and H.A Simon. 1985. Why are some problems hard? evidence from tower of hanoi. *Cognitive Psychology*, 17(2):248–294.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. *Preprint*, arXiv:2306.14050.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. *Preprint*, arXiv:2210.06726.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. *Preprint*, arXiv:2212.08410.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics.

Robert G Miller. 1962. Statistical prediction by discriminant analysis. In *Statistical prediction by discriminant analysis*, pages 1–54. Springer.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *Preprint*, arXiv:2311.11045.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

Allan H Murphy. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. *Preprint*, arXiv:1910.14599.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *Preprint*, arXiv:2112.00114.

Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. Creak: A dataset for commonsense reasoning over entity knowledge. *Preprint*, arXiv:2109.01653.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

TN Palmer, FJ Doblas-Reyes, Antje Weisheimer, and MJ Rodwell. 2008. Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bulletin of the American Meteorological Society*, 89(4):459–470.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *Preprint*, arXiv:2103.07191.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. *Preprint*, arXiv:1808.09121.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. *Preprint*, arXiv:2212.00193.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge. *Preprint*, arXiv:1612.03975.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Preprint*, arXiv:1811.00937.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems. *Preprint*, arXiv:1905.00537.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2023a. Pinto: Faithful language reasoning using prompt-generated rationales. *Preprint*, arXiv:2211.01562.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

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. *Preprint*, arXiv:2201.11903.

Bingsheng Yao, Prithviraj Sen, Lucian Popa, James Hendler, and Dakuo Wang. 2023. Are human explanations always helpful? towards objective evaluation of human natural language explanations. *Preprint*, arXiv:2305.03117.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *Preprint*, arXiv:1904.01130.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795, Singapore. Association for Computational Linguistics.

A Further Discussion

A.1 Intrinsic Mechanism of Impacts of Rationales

In this section, we provide analysis and discussions that take a step further to explain our findings as well as unveil the intrinsic mechanism of rationales.

Why RAFT Causes Performance Drop? We measure the information gain of a rationale as follows. Given a question x, rationale r and answer y, in time step t we construct an input $(x; r_{i<t})$, where $r_{i<t}$ are first i tokens from the rationale. Then we query the model to generate 100 answers using this input. If y appears n times, we record $n/100$ as the probability of the model generating answer y. We repeat such step until the whole rationale is included.

Fig. 7 shows how the probability of the final answer changes with the reasoning process. It fluctuates most time with only one sharp turn that leads to the final answer. This contradicts our expectation that rationales should decompose the task step-by-step, gradually bringing information gain and reducing uncertainty. This observation might be attributed to the inherent differences between human thoughts and LM architecture, it's difficult for LMs to fully mimic human thoughts when their capability is limited in an auto-regressive structure. As a result, it remains to be explored what is the golden rationale for an LM and how to construct it.

Why RAFT Benefit Model Calibration? Previous study attributes the harmful effect of finetuning on calibration to the optimization process. In the optimization process, even when model predictions have already been correct, the loss can be further minimized by increasing the confidence of predictions (Guo et al., 2017), which causes models to be over-confident and less calibrated. We suppose introducing rationales extends the labels from single numbers or letters to a longer text sequence space, which hinders the training process from minimizing loss by raising the confidence in final answers improperly. Besides, a widely used approach in RAFT is to include multiple rationales for one sample, which further strengthens such constraints.

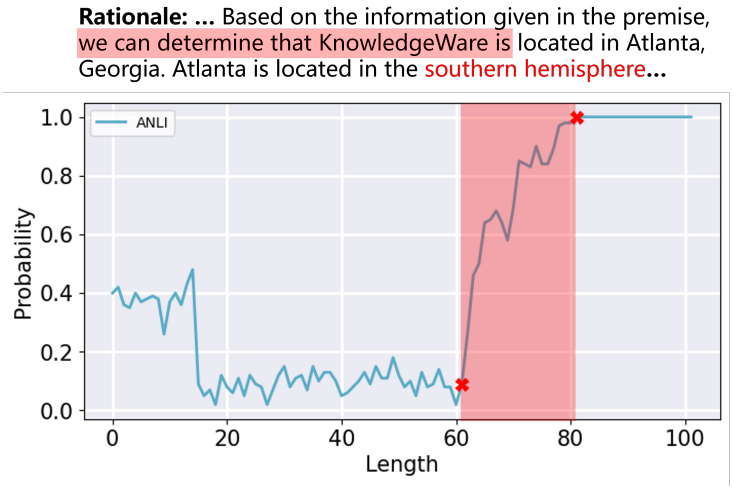

Figure 7: Probability of generating the final answer given different lengths of rationale tokens. Tokens corresponds to the sharp increase are highlighted in red.

A.2 Possible Impacts on General LM Alignment

Though this work mainly studies task-specific finetuning, such conclusion might be generalized to a broader scenario, the alignment of large language models. The most applied technique for alignment for large language models is supervised finetuning (SFT), where human or model responses for a large amount of prompts are collected and used for finetuning models. In typical practice of SFT, each piece of data contains rationales, regardless of the task to which it pertains. At present it is rather unexplored whether we need rationale for all these prompts or how detailed rationales are proper for each prompt.

According to our findings, rationales bring considerable income for some difficult tasks while helping less or distracting models for relatively simpler ones. Re-arranging the rationales in SFT data might enhance the instruction following ability of supervised finetuned models. An intuitive attempt is enhancing, impairing, or even removing the rationales for different prompts according to the task difficulty, which may serve as a refinement and denoising for SFT data. As it is not the main focus of this paper, we leave such exploration for future work.

B Details of Datasets and Licences.

B.1 Introduction of Datasets

Here we introduce details about the datasets we used in the experiments. Note that for test sets whose gold answers are unavailable, we use the validation set instead.

Math Reasoning. MultiArith (Roy and Roth, 2015) is a collection of complicated arithmetic problems designed to test machine learning models. ASDiv (Miao et al., 2020) is an elementary-school-level math word problem corpus that focuses on diversity when constructed. SVAMP (Patel et al., 2021) pays more attention to harder variations of basic problems as most benchmarks focus on difficult problems while models still lack the capabilities to deal with simpler ones. Lastly, GSM8K (Cobbe et al., 2021) is a collection of grade school math problems aimed at challenging the most advanced problem solvers.

Common Sense Reasoning. ARC (AI2Reasoning Challenge) (Clark et al., 2018) and CommonSenseQA (Talmor et al., 2019) are two common sense question answering datasets consisting of science exam questions and human written questions based on concepts in a common sense knowledge-base, CONCEPTNET (Speer et al., 2018). CREAK (Onoe et al., 2021) is a dataset combining common sense and entity knowledge, which includes inference and fact-checking questions for real-world or fictional entities (e.g. Harry Potter).

Sentiment/Topic Analysis. CR (Ding et al., 2008) are SST-2 (Socher et al., 2013) are the two datasets used for sentiment analysis, where each sample is labeled positive or negative. TREC (Li and Roth, 2002) and AGNews (Zhang et al., 2015) are topic classification datasets consisting of questions and news snippets respectively.

Paraphrase. PAWS (Zhang et al., 2019) and QQP[†] consists of sentence pairs from Wikipedia and Quora, which one should determine whether the sentences in the pair have the same semantic meaning.

Natural Language Inference (NLI). CommitmentBank is a natural inference dataset from the SuperGLUE benchmark (Wang et al., 2020) and AdversarialNLI (Nie et al., 2020) is an NLI dataset made up of adversarially constructed questions.

Word Sense Disambiguation. WiC (Pilehvar and Camacho-Collados, 2019) is another sub-task of SuperGLUE, where one should determine whether a given word shows the same meaning in two sentences.

Coreference Resolution. We choose Winogrande (Sakaguchi et al., 2019) in this task, where the task is to identify the specific entity a pronoun points to.

Additional Datasets for Validation SUBJ (Pang and Lee, 2004) is a review analysis dataset, which mainly focuses on deciding the objectivity and subjectivity of reviews for films. CoinFlips (DrewWham and Nascimento, 2020) is a reasoning task where one should decide whether a coin is head or tail after a series of possible flips.

B.2 Licenses of Data and Models

We list all licenses of used data and models here except those whose licenses are not specified. We also cite all original papers for these assets.

[†] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

B.2.1 MIT

SVAMP, GSM8K, CommonsenseQA, CREAK, CR, CoinFlips.

B.2.2 CC

ASDiv (CC BY-NC 4.0), ARC (CC BY-SA), ANLI (CC BY-NC 4.0), WiC (CC BY-NC 4.0).

B.2.3 Apache-2.0

Winogrande, Qwen models

B.2.4 Custom Licenses

AGNews (non-commercial use[‡])
PAWS (free use for any purpose[§])
QQP (Quora ToS, non-commercial use[¶])
CommitmentBank (SuperGlue, research-only[‖])
LLaMA2 (Llama 2 Community License Agreement[**])

B.2.5 Unknown

MultiArith, SST2, TREC, SUBJ

[‡] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
[§] https://github.com/google-research-datasets/paws/blob/master/LICENSE
[¶] https://www.quora.com/about/tos
[‖] https://super.gluebenchmark.com/faq
[**] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf/blob/main/LICENSE.txt

C Brief Quality Assessment for Generated Rationales

We sample 20 data points for each of the 18 datasets, resulting in a total 360 data points. We prompt GPT-4o with the following instruction to assess the helpfulness of each rationale.

Prompt Templates for Rationales Quality Assessment

Following you will see a question and the corresponding answers along with rationales generated by a language model. You should assess whether the reasoning steps in the rationales are helpful for getting the correct answer. Give your analysis first and conclude in the end with "Helpful" or "Unhelpful".

In the evaluation result, only 1 of the rationales are labeled as "Unhelpful", which is then examined by human and turns out to be incorrectly classified by GPT-4o. Such results also demonstrate the quality of the rationales. Details of the "Unhelpful" sample are shown below:

The False Negative Sample

Question:
Decide whether the sentiment of the given statement is positive or negative: with mediasource being so easy , i found no use for the bundled nomad explorer software

Answer:
Let's think step by step.
(1) Mediasource being easy to use suggests it is a positive aspect, making managing media files simple and efficient.
(2) The statement mentions finding no use for the bundled Nomad Explorer software. This implies that the software was not needed as the mediasource is already easy enough, indicating a negative experience.
Though there are positive parts in the sentence, The sentiment of the statement leans towards being negative."""

GPT-Evaluation:
1. The rationale correctly identifies that the statement describes Mediasource as easy to use, which implies a positive experience with this tool.
2. The statement mentions that the user found n̈o use for the bundled Nomad Explorer software.T̈his does not necessarily indicate a negative experience with Nomad Explorer itself. It could simply mean that Mediasource was sufficient for the user's needs, rendering Nomad Explorer redundant. The rationale incorrectly interprets this as a negative sentiment.
...Unhelpful

Human-Evaluation:
Helpful

D Prompts Used in Rationales Generation.

D.1 Prompt Templates for Rationale Generation

Here are the prompt templates for rationale generation, where red texts are fields from datasets and blue texts are content generated by LLM.

Prompt Templates for Rationales Generation (Annotator A)

[System Prompt]
```
You are a helpful assistant.
```

[Math Reasoning]
```
Solve the following math problems. Add a line "The answer is n" at the end
where n is the answer value.

Question: {question}

Answer: Let's think step by step. {LLM output}
```

[Natural Language Inference]
```
Following is a premise and a hypothesis, determine whether the hypothesis
is entailed by the premise, contradictory with the premise, or can not be
determined. Add a line "The answer is x" in the end where x is your choice.

Premise: {premise}
Hypothesis: {hypothesis}
Options: (A) Entailment (B) Contradiction (C) Neutral

Answer: Let's think step by step. {LLM output}
```

[Sentiment Analysis]
```
Follwing is a statement, determine whether the sentiment is negative or
positive. Add a line "The answer is x" in the end where x is your choice.

Statement: {statement}
Options: (A) positive (B) negative

Answer: Let's think step by step. {LLM output}
```

[Topic Classification (TREC)]
```
Following is a question, determine which topic it is about. Add a line "The
answer is x" in the end where x is your choice.

Question: {question}
Options: (A) Description and abstract concept (B) Entity (C) Abbreviation (D)
Human being (E) Location (F) Numeric value

Answer: Let's think step by step. {LLM output}
```

[Topic Classification (AGNews)]
Following is a piece of news and its brief description, determine which topic it is about. Add a line "The answer is x" in the end where x is your choice.

News: {news}
Options: (A) World (B) Sports (C) Business (D) Science and Technology

Answer: Let's think step by step. {LLM output}

[Common Sense (ARC&CSQA)]
Following is a selective question and its answer options. Select the most possible one. Add a line "The answer is x" in the end where x is your choice.

Question: {question}
Options: {options}

Answer: Let's think step by step. {LLM output}

[Common Sense (CREAK)]
Following is a statement, determine whether is is true or false based on common sense and fact. Add a line "The answer is x" in the end where x is your choice.

Statement: {statement}
Options: (A) False (B) True

Answer: Let's think step by step. {LLM output}

[Paraphrase]
Following are two similar sentences. Determine whether they are asking the same question or describing the same situation. Add a line "The answer is x" in the end where x is your choice.

Sentence1: {sentence1}
Sentence2: {sentence2}
Options: (A) Different (B) Same

Answer: Let's think step by step. {LLM output}

[Coreference Resolution]
Following is a sentence where a word is replaced with a blank symbol "_". You will be given two options and you should choose the most possible one to fill in the blank. Add a line "The answer is x" in the end where x is your choice.

Sentence: {sentence}
Options: {options}

Answer: Let's think step by step. {LLM output}
[**Word Sense Disambiguation**]
Following is a target word and two sentences. Determine whether the words in the two sentences have the same semantic meaning. Add a line "The answer is x" in the end where x is your choice.

Target word: {target word}
Sentence1: {sentence1}
Sentence2: {sentence2}
Options: (A) Different (B) Same

Answer: Let's think step by step. {LLM output}

D.2 Prompt Templates for Different Lengths

Prompt Templates for Different Lengths

[**1**]...Give a thorough analysis of the problem and explain your solution.

[**2**]...Analyze the problems and explain your solution as detailed as possible.

[**3**]...Let's think step by step.

[**4**]...Explain your solution with a few words.

[**5**]...Explain the solution as short as you can.

D.3 Prompt Templates from Annotator B

Prompt Templates for Rationales Generation (Annotator B)

[**Natural Language Inference**]
You are now required to perform a natural language inference task. I will provide you with a premise and a hypothesis, and you need to determine whether the hypothesis can be inferred from the premise.

[**Sentiment Analysis**]
You are now required to perform a sentiment analysis task. I will give you a text passage, and you need to determine whether it is positive or negative.

[**Topic Classification**]
You are now required to perform a topic classification task. I will give you a sentence, and you need to determine which of the six possible topics it belongs to.

[**Common Sense (ARC)**]
You are now required to perform a coreference resolution task. I will give you a text passage in which I intentionally omit a word, and then provide you with two options. You need to determine which option is more fitting in the context.

[**Coreference Resolution**]

You are now required to perform a coreference resolution task. I will give you a text passage in which I intentionally omit a word, and then provide you with two options. You need to determine which option is more fitting in the context.

D.4 Prompt Templates from Annotator C

Prompt Templates for Rationales Generation (Annotator C)

[**Natural Language Inference**]
Review a given premise, determine whether the relevant hypothesis can be logically inferred from the premise.

[**Sentiment Analysis**]
You will see a passage of text, please determine the sentiment of the text.

[**Topic Classification**]
Given a sentence, please determine which of the six categories it belongs to.

[**Common Sense (ARC)**]
You will be shown a question and four answers, from which you need to choose one based on your common sense.

[**Coreference Resolution**]
Given a passage of text where a word is left blank, and you will see two options, each corresponding to a word. Please infer, based on the context, which word is more suitable to fill in the blank.

E Training Settings.

Table 2 is the hyper-parameters used in finetuning models. All models are trained for 3 epochs using Huggingface Transformers Library[††] on 4 NVIDIA A-100 GPU and Fully Sharded Data Parallel (FSDP).

Table 2: Hyper-parameters of finetuning.

PARAMETERS	VALUES
EPOCH	3
LEARNING RATE	$5e^{-6}$
BATCH SIZE PER DEVICE	4
GRADIENT ACCUMULATION	4
GRADIENT CHECKPOINTING	TRUE
PRECISION	BF16
MAX LENGTH	2048
WARMUP RATIO	0.03
WEIGHT DECAY	0
LEARNING RATE SCHEDULER	COSINE

Below are an overview of training and inference time.

Table 3: Training/Inference time overview.

MODEL SCALE	TRAINING TIME (MINS) W/O RATIONALES	TRAINING TIME (MINS) W/ RATIONALES
7B	~25	~40
13B	~60	~90
INFERENCE TIME (SECONDS PER SAMPLE)		
7B	~0.12	~40
13B	~3.5	~60

F Correlation between V-usable Information and Lengths of Rationales

Table 4 is the V-usable Information and average length of rationales for a dataset in each task category, where a distinct linear correlation comes up.

Table 4: The V-usable Information and average length of rationales for a dataset in each task category. We denote V-usable Information as V-INFO and average length of rationales as LENGTHS. The Pearson Coefficient is 0.895.

METRICS	SST2	ANLI	PAWS	WIC	WINOGRANDE	ARC CHALLENGE	GSM8K
V-INFO	-0.222	-0.0013	0.0011	0.0023	0.0044	0.0014	0.178
LENGTHS	176.1	202.7	232.2	231.3	233.3	254.8	339.7

[††]https://huggingface.co/docs/transformers/index

G Results of Different Models and Hyper-parameters

G.1 Different Models

Fig. 8 and Table 5 are experimental results of different models and hyper-parameters. We can still observe significant linear correspondence between the improvement in accuracy and ECE in other models with high Pearson Coefficient and near-zero P-value. Besides, when models are trained with different hyper-parameters, our main conclusions still hold. Fig. 9 shows results of LLaMA-2 fine-tuned on rationales generated with LLaMA-3.1-70B-Instruct and GPT-4o-mini. The conclusions are the same.

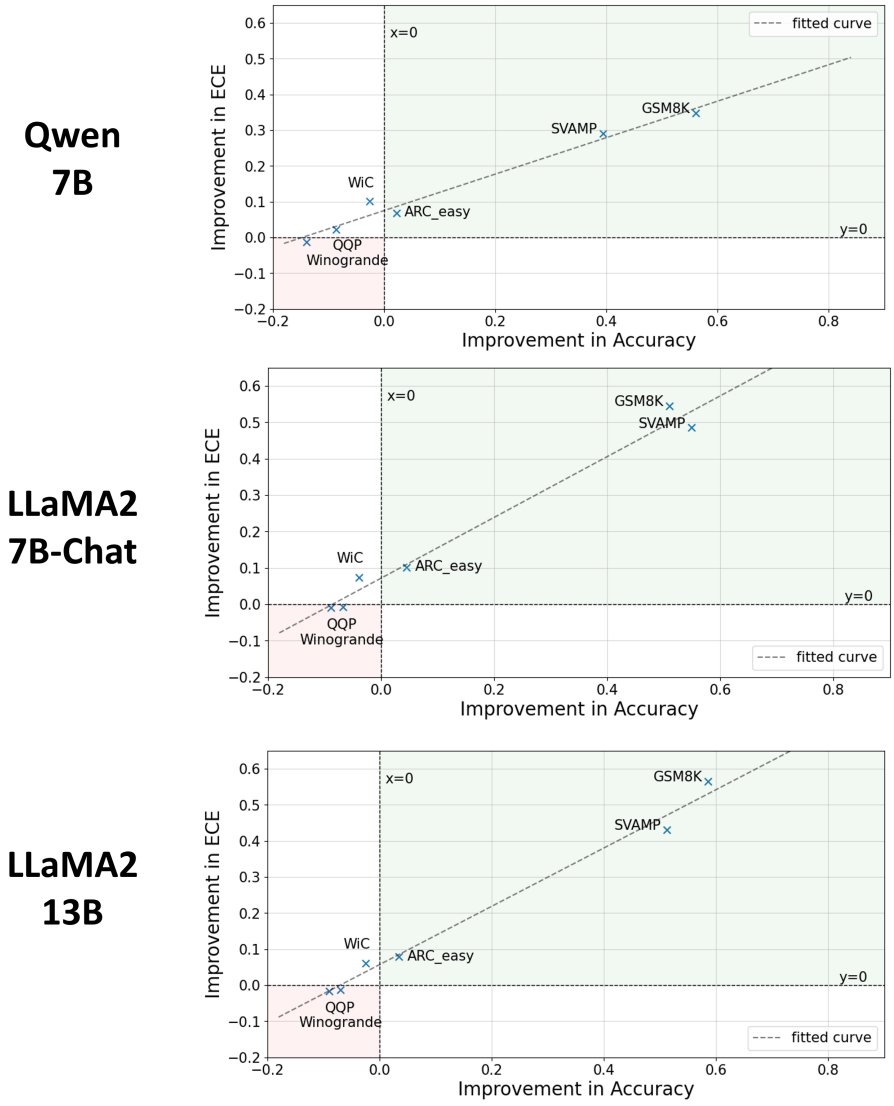

Figure 8: Improvement in Accuracy and ECE of different models. Significant linear correspondence can still be observed.

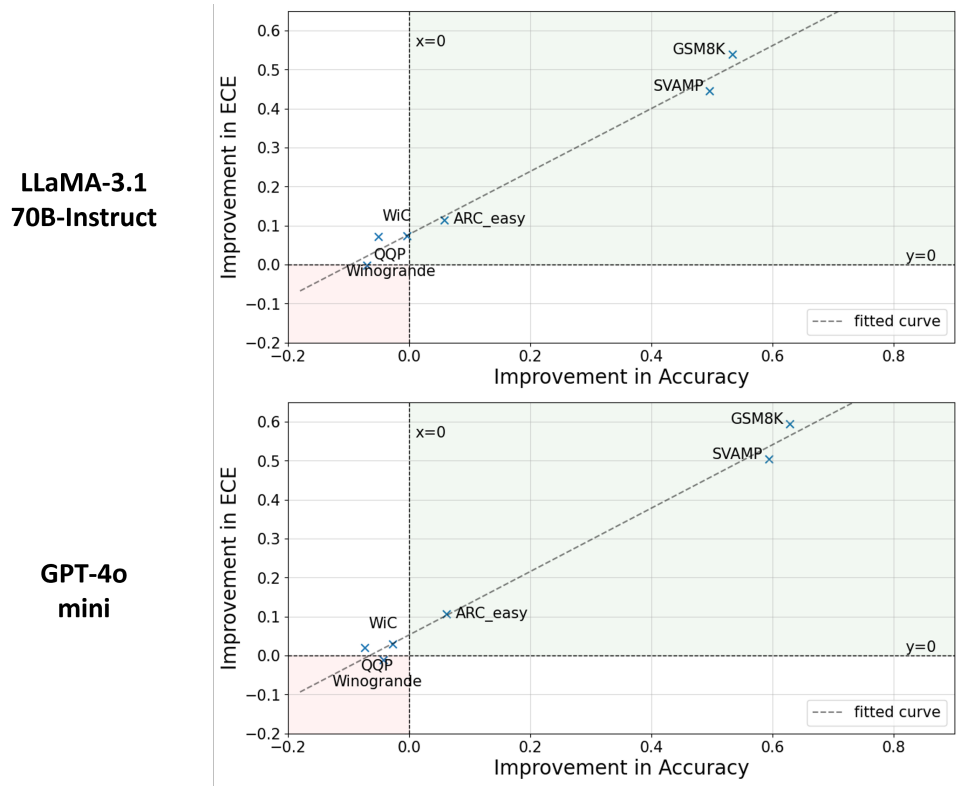

Figure 9: Improvement in Accuracy and ECE of models trained on rationales generated with LLaMA-3.1-70B-Instruct and GPT-4o-mini. Significant linear correspondence can still be observed.

Table 5: Significance test of linear correspondance in different models.

Models	Pearson	P-value
Qwen-7B	0.9883	1.024×10^{-4}
LLaMA2-7B-Chat	0.9902	7.248×10^{-5}
LLaMA2-13B	0.9943	2.469×10^{-5}

G.2 Hyper-parameters

Table. 6 shows the results from the best models in the parameter search. For comparison, we also list results of our main experiments. As can be seen in the table, it remains unchanged whether RAFT brings improvement or harm to model performance and calibration.

Table 6: Results from best models in the parameter search.

Metric	ΔAcc (paper)	ΔAcc (w/ parameter search)	ΔECE (paper)	ΔECE (w/ parameter search)
QQP	-0.092	-0.047	-0.032	-0.004
CR	-0.002	-0.006	0.003	0.01
GSM8K	0.509	0.509	0.544	0.394

H ECE Results Produced with Different Prompts

Following are ECE results of models trained with rationales produced with different prompts. We can see that difference of ECE is minor among models trained with different rationales and our conclusion consistently holds.

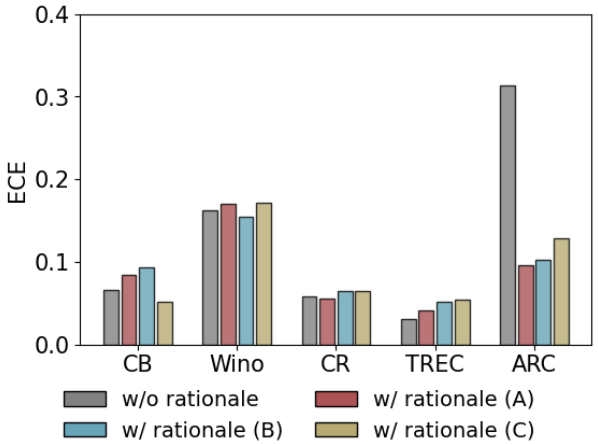

Figure 10: ECE of models trained with rationales produced by annotator B and C.

I Details of Blank Rationales

Here are experimental details of Section 6.1. We conduct experiments on three datasets, WiC, QQP and Winogrande. For each sample, we substitute the rationale with a blank sequence composed of <Think_n> to ensure that it does not provide any additional information, as is shown in Table 8. The number of blank think tokens is set as the average length of rationales in the dataset, specific numbers are listed in Table 7.

Table 7: Length of blank rationales for each dataset.

Datasets	Number of Inserted <Think_n>
WiC	55
QQP	38
Winogrande	32

Table 8: Illustration of blank rationales.

Rationale	The premise provides information about KnowledgeWare, its founders, its headquarters in Atlanta, Georgia, and its product ... The answer is B
Blank Rationale	<Think_0> <Think_1> <Think_2> <Think_3> <Think_4> <Think_5> <Think_6> ... <Think_n> The answer is B

Experimental results are shown in Figure 11

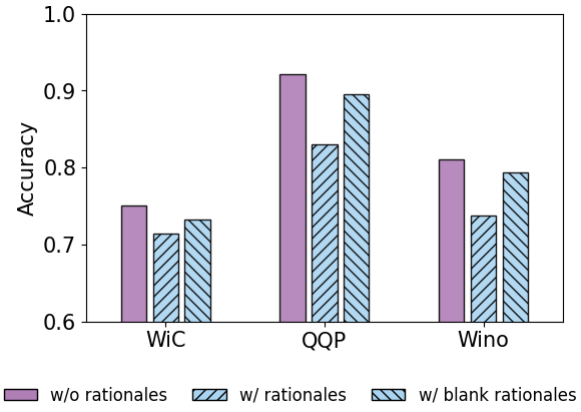

Figure 11: Experimental results of blank rationales.

J Results of Rationale-Augmented Prompting Setting.

Here are the diagrams that supports the conclusions in RAP setting.

Fig. 12 shows the improvement in accuracy and ECE respectively, where the conclusions remain that rationales deteriorate model performance in some tasks, while in most cases benefit model calibration. Fig. 13 and Table 9 shows the relation between improvement in accuracy and ECE. In Fig. 9 we zoom the dense area and omit point labels for clearance. There are still linear correlation between improvements in model accuracy and ECE, while such correlation may be less visually significant as most points are near the origin. Fig. 14 shows the relation between Acc/ECE improvement and different difficulty metrics, in which the positive relation is less pronounced than that in RAFT experiments, which indicates that in RAP setting, model performance and calibration are less affected by task difficulty.

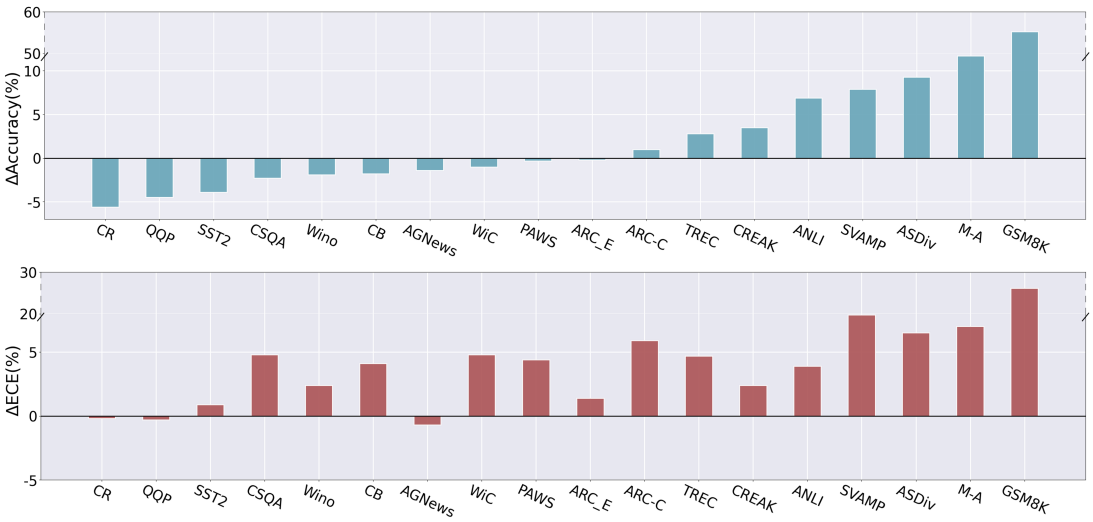

Figure 12: Improvements in accuracy and ECE under rationale-augmented prompting setting. Datasets are re-ordered according to the improvements in accuracy. Wino refers to Winogrande and M-A refers to MultiArith.

Table 9: Numeric value of the linear fit between Accuracy and ECE Improvment.

RESULTS	VALUES
SLOPE	0.413597
INTERCEPT	0.030452
PEARSON	0.955723
P-VALUE	$3.230854e^{-10}$

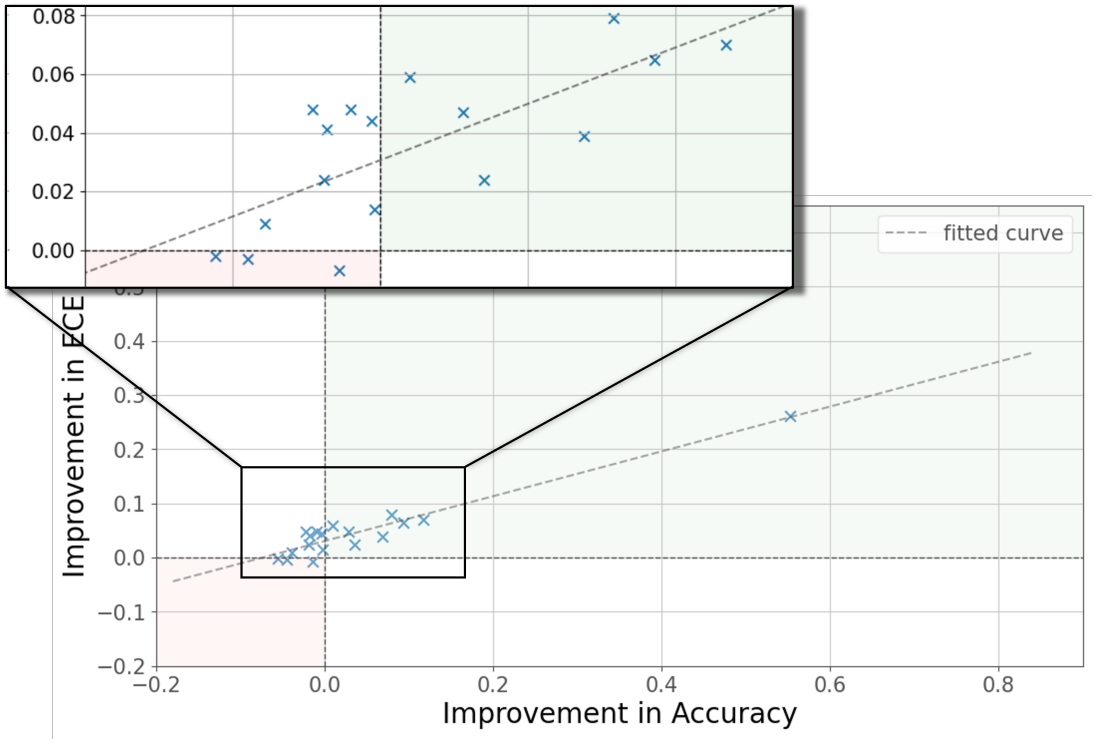

Figure 13: Improvement in Accuracy and ECE of `gpt-3.5-turbo-0613` under rationale-augmented prompting setting. Each point is a datasets, and its x/y-coordinate represents the improvement in model accuracy/ECE respectively. Point labels are omitted as points are close to each other.

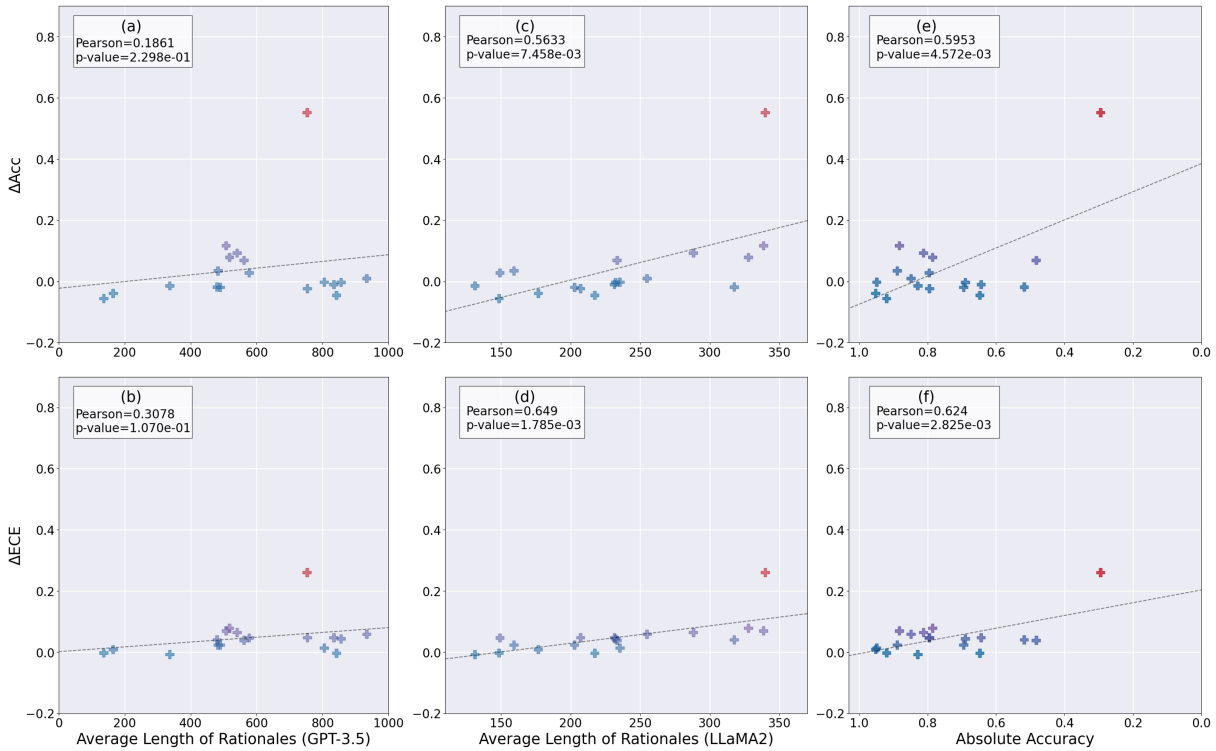

Figure 14: Improvement in accuracy and ECE under rationale-augmented prompting setting when task difficulty is measured with different metrics. Subplot (a), (b): Average length of rationales generated by `gpt-3.5-turbo-0613`. Subplot (c), (d): Average length of rationales generated by LLaMA-2-7B-base. Subplot (e), (f): Original Accuracy of LLaMA-2 model finetuned with answer labels (x-ticks is reversed).

Table 10: Pearson and p-Value of the one-tailed hypothesis test for linear relation. Bold entries are p-values lower than 0.05.

Settings	Metrics	Accuracy		ECE	
		Pearson	p-Value	Pearson	p-Value
RAFT	GPT-3.5	0.0010	0.4984	0.1161	0.3232
	LLaMA2	0.7059	$\mathbf{5.3e^{-4}}$	0.7802	$\mathbf{6.7e^{-5}}$
	Accuracy	0.9295	$\mathbf{1.2e^{-8}}$	0.9718	$\mathbf{9.3e^{-12}}$
RAP	GPT-3.5	0.1861	0.2298	0.3078	0.1070
	LLaMA2	0.5633	**0.0075**	0.649	**0.0018**
	Accuracy	0.5953	**0.0046**	0.6240	**0.0028**

K Results of Multi-task Training

Fig. 15 shows the improvement in accuracy and ECE of models trained in multi-task settings. In polarity mixture setting, for data where RAFT brings gain, we mix QQP, Winogrande, and CR. For the other one, we mix GSM8K, ARC, and CREAK. All results under multi-task learning setting are on the same side of y-axis with baselines, which indicates that multi-task instruction tuning does not change whether RAFT acts positively or negatively.

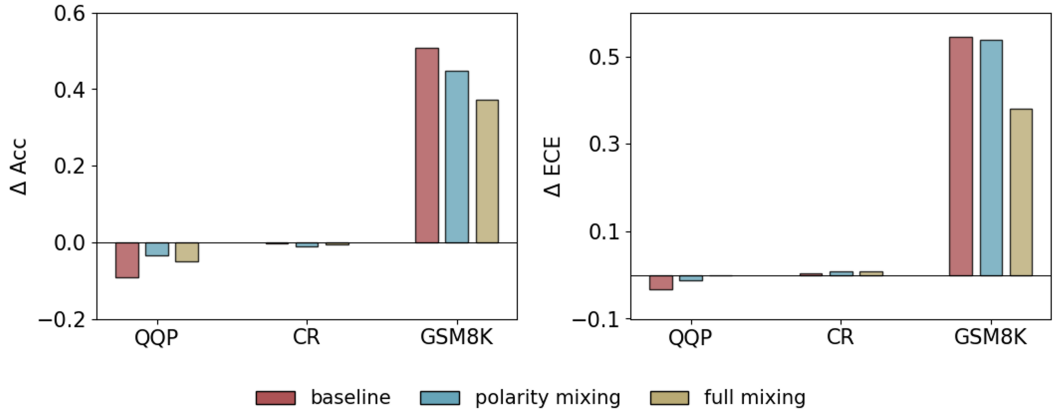

Figure 15: Improvement in Accuracy and ECE of models trained in multi-task settings. (a) Baseline: single task finetuning. (b) Polarity mixture: mix data from datasets where RAFT cause performance increase (or drop). (c) Full mixture: mix data from all datasets.