ATMANRL: TOWARDS FAITHFUL REASONING VIA DIFFERENTIABLE ATTENTION SALIENCY

Max Henning Höth
Aleph Alpha Research
Lab1141

Kristian Kersting
TU Darmstadt
Hessian.AI
Lab1141

Björn Deiseroth
Aleph Alpha Research
Lab1141

Letitia Parcalabescu
Aleph Alpha Research
Lab1141

ABSTRACT

Large language models (LLMs) increasingly rely on chain-of-thought (CoT) reasoning to solve complex tasks. Yet ensuring that the reasoning trace both contributes to and faithfully reflects the processes underlying the model's final answer, rather than merely accompanying it, remains challenging. We introduce ATMANRL, a method that leverages differentiable attention manipulation to learn more faithful reasoning through reinforcement learning. By training an additive attention mask that identifies tokens in the CoT crucial for producing correct answers, we derive a saliency reward signal that encourages the model to generate reasoning traces that genuinely influence its final predictions. We integrate this saliency reward with outcome-based rewards within the GRPO framework to jointly optimize for correctness and interpretability. Experiments on GSM8K with Llama-3.2-3B-Instruct demonstrate that our approach can identify influential reasoning tokens and enable training more transparent reasoning models.

1 INTRODUCTION

Chain-of-thought (CoT) prompting (Wei et al., 2022), supervised learning and reinforcement learning (RL) approaches (Yang et al., 2025; OpenAI et al., 2024; Guo et al., 2025) eliciting reasoning traces have improved the reasoning abilities of large language models (LLMs). By generating intermediate reasoning steps before the final answer, models often reach higher accuracy on complex tasks.

The presence of a reasoning trace, however, does not guarantee that the model actually uses it to arrive at its answer. Consequently, a central question is: *Does the generated CoT causally influence the model's final prediction and has explanatory power, or does it merely accompany it as a stylistic artifact?* This question relates to the notion of *faithfulness*, which asks whether an explanation reflects the model's true decision-making process (Jacovi & Goldberg, 2020). An unfaithful reasoning trace may appear *plausible* and logically coherent while the model reaches the correct answer through shortcuts that bypass the stated reasoning (Agarwal et al., 2024). Prior work shows that LLMs can produce plausible-sounding CoT explanations that do not align with the mechanisms that drive their predictions (Turpin et al., 2023; Lanham et al., 2023; Barez et al., 2025).

To investigate this gap, we distinguish between *saliency* and *faithfulness*. We define saliency as the measurable causal contribution of individual reasoning tokens to the final answer logits. Faithfulness requires more, namely, the reasoning trace must accurately reflect the latent reasoning that produces the answer. Saliency, therefore, constitutes a necessary but not sufficient condition for faithfulness. Ensuring *saliency of the reasoning trace*, defined as the measurable influence of reasoning tokens on the final prediction, prevents CoT from degenerating into lengthy yet weakly relevant narratives. Without such constraints, reasoning traces risk functioning as post-hoc rationalizations rather than interpretable evidence of the model's computation.

Guided by this distinction, we propose ATMANRL to enforce reasoning trace saliency, a method that explicitly trains models to produce salient reasoning traces using reinforcement learning. Our approach builds on ATMAN (Deiseroth et al., 2023), an attention manipulation technique that allows targeted modification of attention weights through a predefined mask. Whereas prior work uses ATMAN for post-hoc interpretability, we instead treat the attention manipulation mask as a learnable,

differentiable object. This allows us to: (i) *efficiently* identify which tokens in the reasoning trace are truly influential for the final answer, (ii) derive a saliency-based reward signal from these contributions, and (iii) incorporate this signal into reinforcement learning to encourage the generation of salient reasoning steps while discouraging extraneous or weakly relevant explanatory content.

Overall, our contributions are: (1) We introduce a **saliency reward** derived from optimizing a differentiable attention that identify salient tokens in the CoT. (2) We combine this saliency reward with outcome-based rewards in the GRPO framework to **jointly train for correctness and reasoning quality** in terms of saliency. (3) We evaluate our method on GSM8K using Llama-3.2-3B-Instruct and show that we can *reduce extraneous reasoning while preserving accuracy*.

2 Related Work

CoT / Reasoning Traces. CoT prompting (Wei et al., 2022) and RL methods such as GRPO (Shao et al., 2024) encourage LLMs to generate reasoning traces. RL improves reasoning performance by optimizing outcome-based rewards. However, outcome rewards focus on answer correctness and do not enforce that the reasoning trace causally influences the final prediction. In contrast, we explicitly reward causal dependency between CoT tokens and the answer.

Reasoning Trace Faithfulness. The faithfulness of model explanations has been studied extensively in interpretability research. Work demonstrated and argued that CoT explanations can be unfaithful, with models sometimes reaching correct answers through reasoning that contradicts their stated logic (Turpin et al., 2023; Lanham et al., 2023; Parcalabescu & Frank, 2024; Barez et al., 2025). Process reward models assign rewards to intermediate reasoning steps using external supervision (Lightman et al., 2023) and improve the plausibility of CoT. However, plausibility reflects consistency with an external evaluator, not alignment with the model's internal computation. Faithfulness instead reflects the model's mechanisms that causally produce the answer. Therefore in our method, we learn an attention mask for each sample to verify the causal influence of each token.

Critical Reasoning Tokens. Work showed that individual CoT tokens (called *critical tokens*) can play outsized influence on LLM outputs (Lin et al., 2025). Vassoyan et al. (2025) encouraged exploration on such tokens to improve RL fine-tuning efficiency. Yan et al. (2024) intervene on attention weights to mitigate over-reliance on misleading tokens in few-shot examples. Unlike these methods, which analyze or manipulate reasoning tokens post hoc, we use differentiable attention manipulation to learn token-level saliency and incorporate it into RL training.

Attention Manipulation. AtMan (Deiseroth et al., 2023) introduced memory-efficient attention manipulation for transformer interpretability, enabling targeted suppression of individual tokens' to estimate their influence. We frame AtMan as a differentiable attention mask and optimize it toward correct answers via SGD, to identify salient reasoning tokens.

3 Differentiable attention manipulation for faithful reasoning

In the following, we introduce our method AtManRL to train models to produce salient reasoning traces by framing AtMan as a differentiable attention mask. Specifically, we (1) recap AtMan, (2) describe how we learn the mask, (3) derive a saliency measure from the optimized mask, and finally, (4) integrate saliency as an RL reward during training.

3.1 Background: AtMan Attention Manipulation

First, we review the additive AtMan-attention manipulation introduced in Deiseroth et al. (2023). In a standard transformer, attention outputs are computed as: $O = \mathrm{softmax}(H) \cdot V$, where $\cdot$ denotes matrix multiplication and where where the pre-softmax attention scores are given by $H = QK^T/\sqrt{d}$. Here, $Q, K, V \in \mathbb{R}^{h \times s \times d}$ denote the query, key, and value tensors with h attention heads, sequence length s, and head dimension d. Atman manipulates the pre-softmax scores H with an additive mask $H^{\text{AtMan}} \in \mathbb{R}^{s \times s}$:

$$H = Q \cdot K^T/\sqrt{d} + H^{\text{AtMan}} \tag{1}$$

Applying the mask H^{AtMan} before the softmax ensures that the resulting attention scores still add to one after the softmax. Additionally, unlike other perturbation methods in XAI (e.g. Shapley

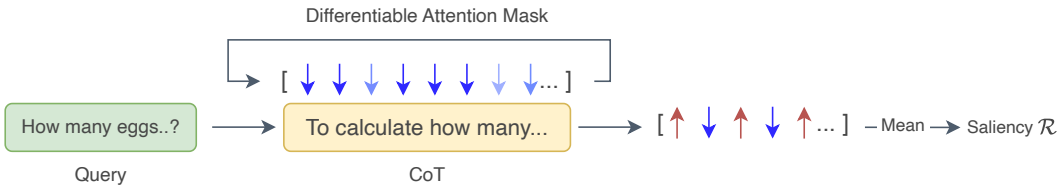

Figure 1: We initialize the additive attention mask H^{AtMan} with a negative value to suppress attention over CoT tokens. We then optimize the mask for 200 steps to restore the correct answer probability.

values), this does not introduce a shift in the input-distribution or positional embeddings, but carefully manipulates the attention of the model of every single token. Positive mask values increase attention to selected tokens, whereas negative values suppress their influence. For autoregressive models, we additionally apply a lower-triangular causal mask T and compute $H_M = H \circ T$, where $\circ$ denotes the Hadamard product. Deiseroth et al. (2023) used H^{AtMan} to suppress the attention to individual tokens by assigning a fixed negative value – treated as a hyperparameter – to the corresponding columns of H^{AtMan} to analyse each individual token's impact on the output logits of the LLM.

3.2 Training an H^{AtMan} mask for Measuring Saliency

Because the mask enters the pre-softmax attention scores additively, it remains fully differentiable. We restrict H^{AtMan} to tokens within the reasoning trace (CoT) and do not modify attention over prompt tokens or final answer tokens. The prompt remains fixed and outside the model's control, and therefore does not constitute a target for reward shaping. Conversely, we require that the final answer depends causally on the reasoning trace. If the reasoning trace is salient, perturbing its attention should affect the probability of the correct answer.

We initialize all CoT-related mask entries with a negative constant $c = -0.4$. This initialization uniformly suppresses attention to reasoning tokens and produces a flatter post-softmax distribution. From this suppressed state, we optimize the mask to restore the probability of the correct answer.

Specifically, to train the mask, we minimize the cross-entropy loss of the ground-truth answer tokens $y_{1:N}$ under teacher forcing:

$$\mathcal{L}_{\text{mask}} = -\frac{1}{N}\sum_{n=1}^{N} \log P\big(y_n \mid c_{1:T}, y_{1:n-1}, H^{\text{AtMan}}\big), \tag{2}$$

where $c_{1:T}$ denotes the CoT tokens. The mask is the only trainable object at this stage to identify attention configurations that preserve answer likelihood under suppressed reasoning.

We stop optimizatizing the mask after a fixed number of steps. We normalize the mask by dividing by the initialization constant $\hat{H}^{\text{AtMan}} = H^{\text{AtMan}}/c$ and compute the average normalized mask value over the lower-triangular (causal) region:

$$\mathcal{R}_{\text{Faithfulness}}(a_i) = \frac{1}{|\mathcal{I}_v|}\sum_{w\in\mathcal{I}_v} \hat{H}^{\text{AtMan}}_{w,v}, \quad \mathcal{I}_v = \{w \in \{1,\dots,n\} \mid w \geq v\}. \tag{3}$$

This quantity serves as our saliency measure and reward for rollout a_i. Intuitively, it measures how strongly the reasoning tokens must be re-enabled to preserve the correct answer probability.

3.3 Optimizing Saliency via Reinforcement Learning (RL)

For RL, we combine the saliency reward with a standard outcome reward:

$$\mathcal{R}_{\text{Outcome}}(a_i) = \begin{cases} 0 & \text{if } i = j, \\ -1 & \text{otherwise}, \end{cases} \tag{4}$$

where j denotes the ground-truth answer and a_i the rollout prediction.

Thus, the **total reward** is $\mathcal{R}_{\text{total}}(a_i) = \mathcal{R}_{\text{Outcome}}(a_i) + \mathcal{R}_{\text{Faithfulness}}(a_i)$. Following GRPO (Shao et al., 2024), we compute the group-normalized reward $\hat{\mathcal{R}}_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{R}_{\text{total}}(a_i)$, and define the

	Avg. token / CoT ↓	Numbers % ↑	Stop words % ↓	Symbols % ↑	Pass@4 ↑
AtManRL	**119.7** (-36%)	**15.94** (+33%)	**25.23** (-14%)	**8.45** (+46%)	89.4 (-2%)
Baseline	187.8	11.98	29.27	5.79	**91.4**

Table 1: Comparison between baseline and ATMANRL on GSM8K. ATMANRL reduces average reasoning length while maintaining comparable performance. Token composition shifts toward information-dense content, with fewer stop words (most common words of a language which usually carry little to no meaning) and more numbers and symbols.

advantage $\mathcal{A}(a_i) = \mathcal{R}_{\text{total}}(a_i) - \hat{\mathcal{R}}_{\text{total}}$. We then update the policy using the clipped GRPO objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \min\left(\frac{\pi_\theta(a_i \mid q)}{\pi_{\theta_{\text{old}}}(a_i \mid q)}\mathcal{A}(a_i),\ \text{clip}\left(\frac{\pi_\theta(a_i \mid q)}{\pi_{\theta_{\text{old}}}(a_i \mid q)}, 1-\epsilon, 1+\epsilon\right)\mathcal{A}(a_i)\right). \quad (5)$$

3.4 Implementation Details

Mask Optimization: We use AdamW (Loshchilov & Hutter, 2019) with a learning rate of $1e-3$, betas of 0.6 and 0.9999 and a weight decay of 0.05. We train for 200 gradient steps to update H^{AtMan}.

Value Scaling: Before adding H^{AtMan} to the attention scores, we clamp to an upper bound of 0 to prevent applying positive values to each token since we just want to detect non salient ones. We then scale them with a factor 10 for faster convergence.

RL Training Details: We fine-tune the model for 5 epochs using GRPO with 8 rollouts per query and a maximum generation length of 1024 tokens. For each update, we compute the saliency reward over batches of 8 queries. We perform two gradient passes per batch using a mini-batch of 2 to recompute policy log-probabilities for the clipped RL objective. We use a fixed 8×10^{-7} learning rate and $\epsilon = 0.2$ standard clipping parameter. We conducted all experiments on 48 NVIDIA A100 GPUs.

4 Experimental Results

We evaluated ATMANRL on 500 randomly sampled GSM8K (Cobbe et al., 2021) problems using `Llama-3.2-3B-Instruct` (Grattafiori et al., 2024). We compare against a *baseline* trained from the same model under identical settings, but optimized solely with the outcome reward.

For each problem, we prompted the model to produce a chain-of-thought (CoT) followed by the final answer, separated by a #### delimiter. We extracted the reasoning trace as all tokens generated before the delimiter and the final answer as the tokens following it. For each query, we sampled four independent reasoning traces (pass@4 evaluation). We computed *answer correctness* from the extracted final answer string. To investigate *changes in reasoning composition*, we tokenized each generated CoT and annotated tokens using *spaCy*'s part-of-speech tagger. We then computed the average reasoning length and the relative frequency of numbers, stop words, and symbols to quantify how the saliency reward affects both reasoning length and token-level composition.

The results summarized in Table 1 show that ATMANRL reduces the amount of tokens per response by 36%, thereby lowering inference cost and eliminating non-salient reasoning content. There is a 4% decrease in stop words and a 7% increase in numbers and symbols, while pass@4 drops by only 2%. The reduction in function words and increase in numbers and symbols align with the mathematical nature of GSM8K, favoring concise, computation-focused reasoning over verbose explanations. These results suggest that the saliency reward encourages the model to retain information-dense tokens that contribute to the final prediction while suppressing tokens with limited causal impact. Evaluation results in App. A.2 and examples in App. A.3.

5 Conclusions

We introduced ATMANRL for training more salient reasoning in LLMs through differentiable attention manipulation. By learning attention masks that identify reasoning tokens crucial for correct answer generation, we derive a saliency reward that we combine with outcome-based rewards in RL training. Our approach provides a principled way to encourage models to produce reasoning traces

that genuinely influence their predictions, rather than plausible but non-causal rationalizations, which should in future experiments also prove the increased faithfulness. Initial experiments on GSM8K demonstrate the feasibility of learning such saliency signals and that we can reduce extraneous reasoning while preserving accuracy when training with such saliency rewards. We discuss limitations and future work in A.1.

REFERENCES

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

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. Chain-of-thought is not explainability. *Preprint, alphaXiv*, pp. v1, 2025.

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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Björn Deiseroth, Mayukh Deb, Samuel Weinbach, Manuel Brack, Patrick Schramowski, and Kristian Kersting. ATMAN: Understanding Transformer Predictions Through Memory Efficient Attention Manipulation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 63437–63460. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c83bc020a020cdeb966ed10804619664-Paper-Conference.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale,

Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,

Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, 2020. URL https://arxiv.org/abs/2004.03685.

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Lin, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM's Reasoning Capability, January 2025. URL http://arxiv.org/abs/2411.19943. arXiv:2411.19943 [cs].

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally

Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.329. URL https://aclanthology.org/2024.acl-long.329/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=bzs4uPLXvi.

Jean Vassoyan, Nathanaël Beau, and Roman Plaud. Ignore the KL Penalty! Boosting Exploration on Critical Tokens to Enhance RL Fine-Tuning, February 2025. URL http://arxiv.org/abs/2502.06533. Accepted for publication in the Findings of the North American Chapter of the Association for Computational Linguistics (NAACL) 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Shaotian Yan, Chen Shen, Wenxiao Wang, Liang Xie, Junjie Liu, and Jieping Ye. Don't Take Things Out of Context: Attention Intervention for Enhancing Chain-of-Thought Reasoning in Large Language Models. In *The Thirteenth International Conference on Learning Representations*, October 2024. URL https://openreview.net/forum?id=W6yIKliMot.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

A Appendix

A.1 Limitations

Our work has several limitations that suggest directions for future research: first and foremost we need to evaluate our method using more models, more datasets, and different domains to prove the effectiveness and adaptability of the developed method and also analyse the saliency dynamics on larger models. Second, our method optimizes token-level saliency as a proxy for faithfulness. In future work, we will verify whether there is actually an increase in faithfulness using a suiting metric which can prove that the internal reasoning process is better represented in the CoT. Third, while our experiments show a small drop in accuracy, we believe this can be minimized through careful hyperparameter tuning. In particular, adjusting the balance between the saliency and outcome rewards, as well as the mask initialization constant and optimization steps, are promising directions for recovering and potentially surpassing baseline accuracy. Finally, optimizing an AtMan mask for each rollout introduces additional computational overhead compared to standard RL fine-tuning. We selected the initialization constant ($c = -0.4$) and the number of mask optimization steps (200) empirically; systematic ablations and more efficient optimization strategies could further improve practicality.

A.2 Evaluation Plots

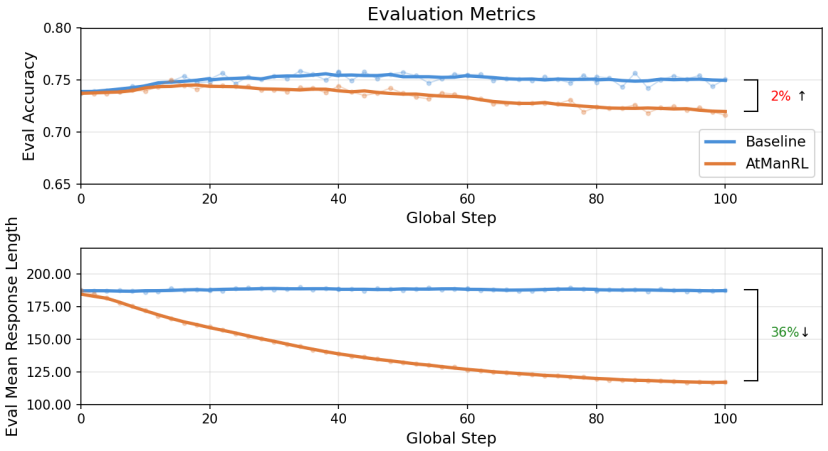

Figure 2: Accuracy and mean response length of evaluation set during training.

A.3 CoT Examples

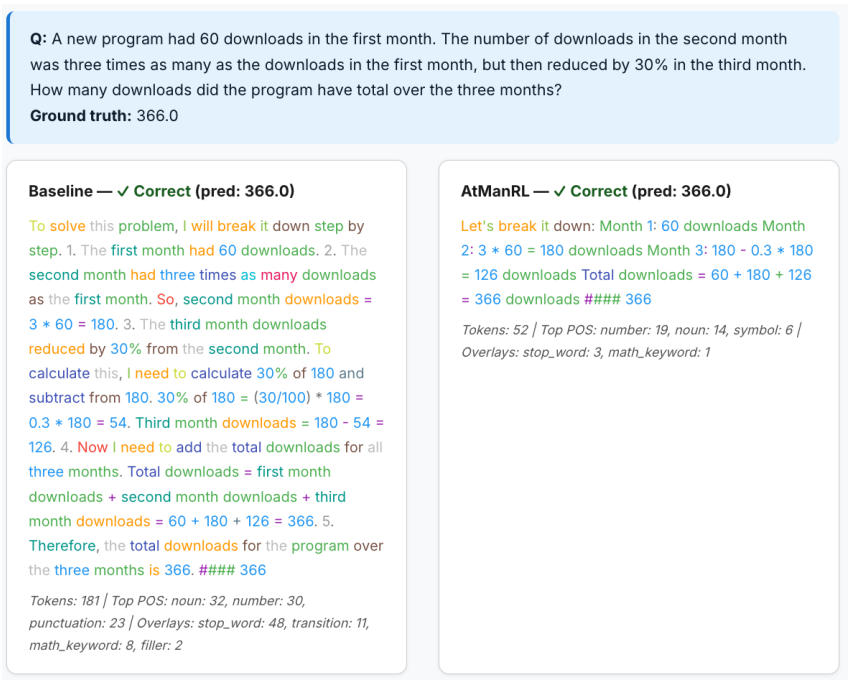

Figure 3: Reasoning trace comparison generated by the baseline and AtManRL final checkpoints. The part-of-speech tagging done by *spaCy*.

Figure 4: Reasoning trace comparison generated by the baseline and AtManRL final checkpoints. The part-of-speech tagging done by *spaCy*.

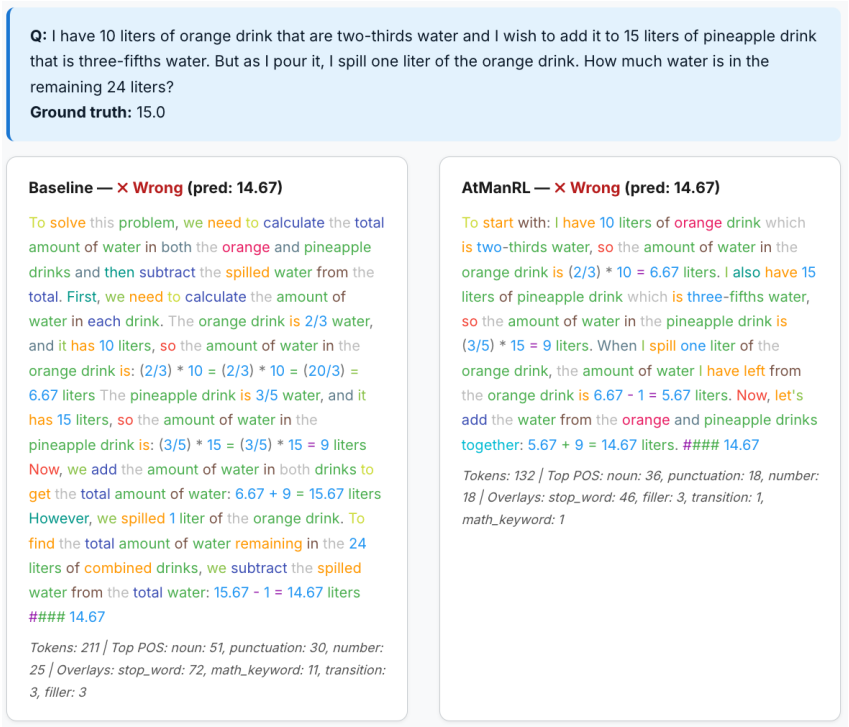

Figure 5: Reasoning trace comparison generated by the baseline and AtManRL final checkpoints. The part-of-speech tagging done by *spaCy*.