Delayed Convergence and Emergent CoT Reliance in RL-Tuned Language Models

Pablo Pérez-Lázaro[1,*], Rocío Aznar-Gimeno[1], Francisco José Lacueva-Pérez[1], Paula Peña-Larena[1] and Rafael del-Hoyo-Alonso[1]

[1]*Instituto Tecnológico de Aragón (ITA)*

Abstract

While Reinforcement Learning Fine-Tuning (RLFT) enhances the performance of Large Language Models (LLMs) on reasoning benchmarks, the internal mechanisms driving these gains are poorly understood. We investigate these dynamics by proposing a mechanistic interpretability methodology to systematically compare a base Qwen-2.5-14B model against its GRPO-fine-tuned and distilled variants. Specifically, we apply a memory-efficient, long-context Tuned Lens to extract layer-wise internal entropy and divergence from the final output, and Integrated Gradients equipped with the new Reasoning Share-mean length-normalized metric to quantify token-level attributions. Our analysis reveals that intermediate log-probability is an unreliable indicator for reasoning capability; instead, reasoning performance results from a shift of internal confidence allocation where RL fine-tuning delays internal convergence, exhibiting prolonged mid-layer exploration before converging sharply at the final output. Equally important, attribution mapping demonstrates a pronounced reliance on intermediate reasoning steps. Despite optimization exclusively via final-answer rewards, the GRPO model shifts its internal attribution toward its self-generated chain, demonstrating that Chain-of-Thought reliance is an emergent computational mechanism driven by RL, rather than a byproduct of supervised fine-tuning.

Keywords

Explainable AI, Mechanistic Interpretability, Large Language Models, Reinforcement Learning, Feature Attribution, Chain-of-Thought Reasoning, Latent Space Probing

1. Introduction

Reasoning in Large Language Models (LLMs) has experienced rapid progress, evolving from a byproduct of next-token generation to achieving state-of-the-art performance on math, commonsense, and symbolic reasoning benchmarks [1, 2]. This progress has driven leading models such as Google's Gemini 3 Flash and Anthropic's Claude Sonnet 4.6 to incorporate RL-based reasoning. Reinforcement Learning Fine-Tuning (RLFT) improved performance in reasoning models [3, 4, 5], but it has been shown that it enhances existing capabilities in the model without expanding the reasoning boundary [6]. However, the underlying mechanisms driving these gains remain opaque. This lack of structural understanding demonstrates the limitations in global explainability techniques, highlighting a critical gap in current interpretability frameworks.

Currently, reasoning explainability is predominantly addressed by observing Chain-of-Thought (CoT) traces post-hoc [1, 7], or through attribution methods that are applied almost exclusively at the local level [8]. Furthermore, literature attempting to quantify reasoning capabilities typically relies on pass@k or accuracy metrics without capturing the underlying generative process [9]. Similarly, recent findings suggest that Reinforcement Learning (RL) does not instill novel reasoning patterns but rather amplifies existing ones as measured by pass@k [6]. However, these studies do not explain the mechanistic changes within the internal layers that drive this improved pass@1 performance. To the

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.

plazaro@ita.es (P. Pérez-Lázaro); raznar@ita.es (R. Aznar-Gimeno); fjlacueva@ita.es (F. J. Lacueva-Pérez); ppena@ita.es (P. Peña-Larena); rdelhoyo@ita.es (R. del-Hoyo-Alonso)

0009-0006-1604-3787 (P. Pérez-Lázaro); 0000-0003-1415-146X (R. Aznar-Gimeno); 0000-0003-0998-2939 (F. J. Lacueva-Pérez); 0000-0001-5750-6238 (P. Peña-Larena); 0000-0003-2755-5500 (R. del-Hoyo-Alonso)

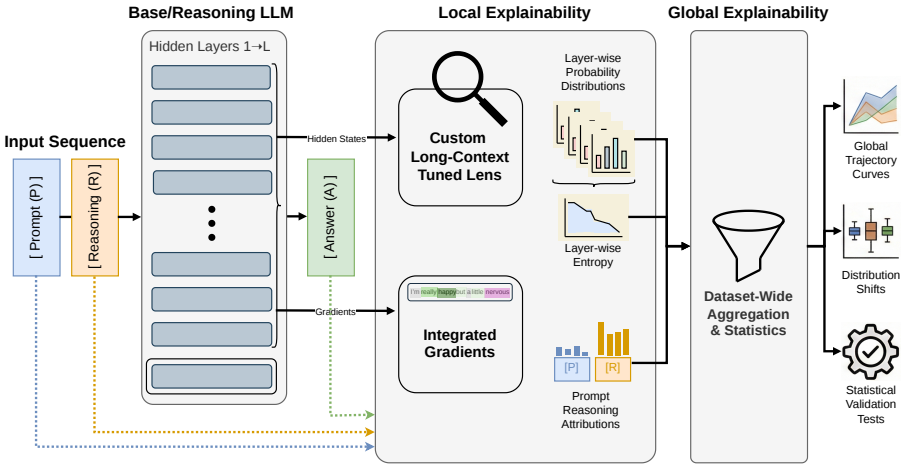


Figure 1: Overview of the proposed mechanistic interpretability pipeline. The framework compares base models with their RL-tuned counterparts by applying a custom Tuned Lens implementation alongside Integrated Gradients to extract layer-wise probabilities and token attributions. By aggregating these local metrics across the entire dataset, the pipeline reveals global structural shifts induced by reinforcement learning.

best of our knowledge, only a limited number of recent studies [10] have identified this gap and begun to explore the intersection between emergent reasoning pathways and internal model dynamics.

We propose an explanation methodology that scales attribution and mechanistic interpretability techniques to facilitate dataset-wide comparisons of reasoning behaviors across models. As illustrated in Figure 1, we first evaluate on a standard reasoning benchmark (GSM8K) using a base Qwen-2.5-14B model and its RL-tuned counterparts. Then, we apply two complementary techniques: a custom long-context Tuned Lens to extract layer-wise probability distributions and internal entropy, alongside Integrated Gradients (IG) to quantify attributions between the prompt and the generated reasoning tokens. Finally, we aggregate these local metrics across the entire dataset, generating global trajectory curves and box plots that elucidate how RLFT alters the model's internal dynamics.

The main contributions of this work are as follows:

- *An interpretability pipeline* that aggregates token-level attributions into aggregate metrics, thereby resolving current limitations in LLM explainability.
- *Open-source engineering advancements*, including a memory-efficient implementation of the Tuned Lens for long-context windows, custom scalar objectives tailored for IG, and a length-normalized metric (RS_{mean}) designed to mitigate token-count bias when evaluating CoT sequences.
- *The identification of distinct internal shifts* in RL-tuned reasoning models, specifically an increase in mid-layer token entropy (indicative of broader internal exploratory search prior to late-stage commitment) and a significant attribution shift wherein models rely substantially more on intermediate reasoning tokens than on the initial prompt.

2. Related Work

LLM explainability is dominated by local post-hoc methods, especially for reasoning tasks where explanations typically trace influential input tokens or intermediate reasoning chains. Mechanistic interpretability probes internal dynamics beyond surface text, but it is still largely local in scope. In LLM reasoning, CoT emerged via prompting with intermediate-step examples, and even simple zero-shot cues (e.g., "Let's think step by step") improved performance on benchmarks such as GSM8K [1, 2]. More recently, work has shifted toward training models for reasoning using RL-style methods, including STaR [3, 4], RLFT with Proximal Policy Optimization (PPO) [5], and Group Relative Policy Optimization

(GRPO), with evidence suggesting RL mostly amplifies reasoning capabilities already accessible through prompting rather than creating them from scratch [6].

To understand the mechanics behind these reasoning gains, post-hoc attribution methods like IG are often employed to assign output contributions to input elements via gradients [11]. Compared to alternatives, SHAP can be prohibitively expensive for high-dimensional LLMs [12], while attention-based heuristics may be unfaithful to the actual decision process [13]. IG has been applied to phenomena such as copying bias in in-context learning [14], motivating its use for analyzing reasoning. While attribution links inputs to outputs, probing-based approaches such as Logit Lens [15] and Tuned Lens [16] decode layer-wise information to inspect evolving predictions. They have been used for trajectory analysis, early exiting, and hallucination-related studies [16, 17].

3. Methodology

We investigate the internal reasoning dynamics of LLMs before and after RLFT. Our framework evaluates these models under zero-shot settings, followed by layer-wise belief extraction using a Tuned Lens and token-level attribution mapping via IG. The complete extraction pipeline is illustrated in Figure 1. We compare three model variants based on the Qwen-2.5 architecture, allowing us to isolate the effects of fine-tuning on latent reasoning:

- *Base Model*: Qwen-2.5-14B (Base). We deliberately select the base version rather than the Instruct variant to prevent latent alignment biases or prior CoT triggers from confounding the zero-shot evaluation.
- *GRPO Fine-Tuned*: The base model fine-tuned using GRPO on the GSM8K train set via the Unsloth framework (1 epoch, LoRA rank/alpha 64/64, learning rate 5e-6). The reward function optimizes strict formatting, XML validation, and exact final-answer correctness.
- *Distilled Model*: DeepSeek-R1 Distill (14B), a Qwen-2.5-14B student supervised on approximately 800k CoT reasoning traces without a secondary RL stage.

3.1. Explainability Techniques

Given an input prompt sequence $x = (x_1, ..., x_n)$, the model autoregressively generates a reasoning span r followed by a numeric answer a. We denote the latent hidden state at transformer layer l for token step t as $h_t^{(l)}$, and the predicted vocabulary distribution at each intermediate layer as $P_t^{(l)}$. We probe the LLM internal dynamics using two complementary mechanistic interpretability techniques adapted for long-context generation.

Layer-wise Probing (Tuned Lens) Because applying a final decoding head to intermediate layers in pre-norm architectures suffers from representation drift, we map $h_t^{(l)}$ to the vocabulary space using a Tuned Lens trained on the validation set from The Pile dataset, applied to all Qwen-2.5 models at a 1-layer granularity.

- *Scalability Contribution*: Standard lens implementations load full trajectory tensors into memory, causing prohibitive memory overhead (exceeding 350 GB RAM) for long reasoning spans (up to 4096 tokens). To manage this memory overhead, we replaced the standard full-trajectory materialization with an iterative, layer-wise pipeline. This custom implementation isolates the hidden states for the target sequence, projects them through the tuned lens, and computes entropy and forward KL statistics directly on the GPU before immediately offloading the results to the CPU and manually flushing the CUDA cache.
- From the decoded distributions, we extract the Shannon Entropy $H(P_t^{(l)})$ to measure the model's confidence at each layer, the log-probability of the target token to track the emergence of the correct prediction, and the Forward KL Divergence $D_{KL}(P_t^{(L)} \parallel P_t^{(l)})$ relative to the final output $P_t^{(L)}$ to quantify representational drift and relative uncertainty regarding the final token prediction. We average these span-wise over both reasoning (r) and answer (a) tokens.

Token Attribution (Integrated Gradients) To quantify LLM reliance on generated Chain-of-Thought (CoT) versus initial prompting, we use Captum's Integrated Gradients (IG). We specifically select IG over alternative attribution methods for two primary reasons: SHAP has proven computationally prohibitive for billion parameter LLMs, and the relationship between raw attention weights and faithful model explanations remains demonstrably unclear. Using a teacher-forcing approach, we apply IG backward from the generated sequence to four scalar targets: the log-probability or entropy of either the first answer token or the full numeric answer a. Log-probability attribution shows which tokens drive the prediction, while entropy attribution reveals which tokens resolve model uncertainty.Global attributions are computed against a zero-vector baseline over 25 integration steps (reduced to 15 for the Distill variant due to longer sequences). To correct for the artificial inflation of absolute sum metrics by lengthy CoT traces, we introduce Reasoning Share-mean (RS-mean). This metric scales the L_1-normalized attribution map, $\text{attr}(t)$, by token-span length to eliminate bias:

$$RS_{\text{mean}} = \frac{\frac{1}{|R|}\|\text{attr}(R)\|_1}{\frac{1}{|P|}\|\text{attr}(P)\|_1 + \frac{1}{|R|}\|\text{attr}(R)\|_1}$$

where R and P denote the subsets of reasoning and prompt tokens, respectively. A higher RS-mean indicates greater functional reliance on intermediate reasoning steps rather than the initial prompt.

3.2. Evaluation Protocol and Metrics

All models were evaluated using the complete test split of the GSM8K dataset (1319 items). Generation proceeded under strict zero-shot settings, employing a minimal `Q: [question]\nA:` template without in-context CoT examples or "Let's think step by step" prompting, and using greedy decoding (temperature 0) to eliminate sampling noise. The final answer was automatically extracted by parsing the right-most integer generated before encountering the specific stop tokens `<|im_end|>` or `\n\n`. Since Tuned Lens metrics aim to assess successful reasoning dynamics, they were computed exclusively on correct answers, while IG metrics were calculated across the full dataset to capture generalized attribution patterns. The significance of differences between model distributions was established via paired Wilcoxon signed-rank tests, with Cliff's δ quantifying effect sizes. All experiments were conducted on a single NVIDIA H100 (80GB).

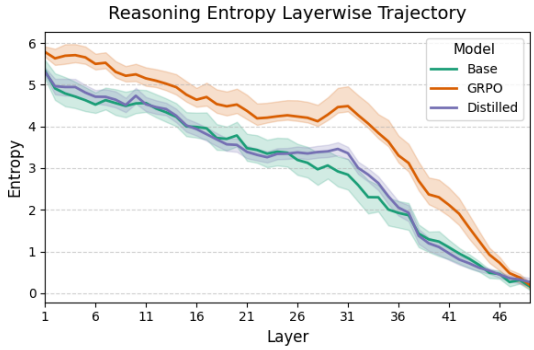
4. Results

Evaluating on the reasoning-intensive GSM8K benchmark confirms the reasoning superiority of RL-aligned models (GRPO: 91.43% (1206/1319), Distilled: 89.61% (1182/1319) vs. Base: 84.31% (1112/1319). While accuracy improved, uncovering the internal cause driving these gains requires deeper analysis.

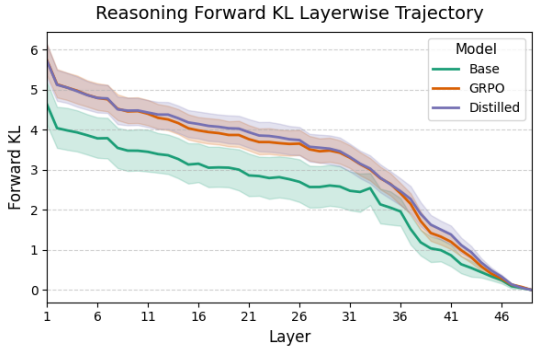
4.1. Internal Dynamics of Reasoning (Tuned Lens)

Our analysis reveals that RL fine-tuning delays internal convergence, exhibiting prolonged exploration in middle layers before forcing a sharp final commitment without altering the log-probability assigned to correct tokens. Figure 2 illustrates the internal belief trajectories across the transformer layers. In contrast to the Base model's rapid convergence toward the final output distribution, both GRPO and Distilled variants sustain elevated entropy and forward KL divergence across intermediate layers (26–38). The elevated entropy and KL divergence in these layers suggest a period of internal search that is absent in the base model, with GRPO yielding a substantially higher total reasoning entropy than the Base model (Wilcoxon $p < 0.001$, Cliff's $\delta = -0.784$). Similarly, the Distilled model exhibits a pronounced delay in representational convergence, as measured by forward KL divergence ($p < 0.001$, $\delta = -0.697$).

While there is intermediate uncertainty, the fine-tuned models become highly decisive at the final layer. Entropy drops sharply to near-zero at the output head (Base vs. GRPO: $p < 0.001$, $\delta = 0.871$),



(a) Layer-wise entropy of the Tuned Lens prediction.



(b) Forward KL divergence between intermediate and final layer distributions.

Figure 2: Internal dynamics across transformer layers (x-axis) averaged over reasoning tokens (CoT steps) for correctly answered GSM8K prompts. Solid lines represent the mean across samples; shaded areas denote the 95% confidence interval. Base model (green) shows rapid convergence, whereas GRPO (orange) and Distilled (purple) maintain high uncertainty and representational drift between layers 26 and 38, indicating prolonged internal search during the reasoning phase.

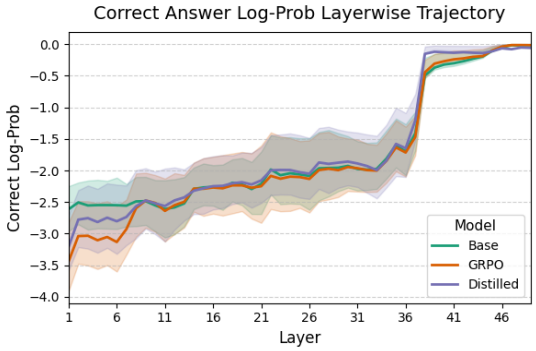


Figure 3: Average log-probability trajectory assigned to the ground-truth final answer token decoded at each transformer layer (x-axis). Solid lines represent the mean log-probability across correctly answered samples, with shaded regions indicating the 95% confidence interval. Despite divergent exploration patterns during the reasoning phase (Fig. 2), all models follow nearly identical representational paths for the final answer token, challenging the use of intermediate log-probability as a direct metric for reasoning quality.

accompanied by a significantly wider log-probability margin between the top-1 and top-2 predictions ($p < 0.001$, $\delta = -0.887$). These results indicate that RL fine-tuning fundamentally alters confidence allocation: it defers commitment during the reasoning process while enforcing strict certainty at the final output.

As illustrated in Figure 3, the layer-wise log-probability assigned to the ground-truth final answer token shows no statistically significant difference among models across correctly answered prompts. This result suggests that intermediate log-probability is an unreliable proxy for reasoning capability, as RL alignment does not enhance the base representational probability of the correct answer within the internal layers. Combined with the entropy and KL-divergence shifts observed during reasoning, these findings indicate that performance gains are strictly procedural, resulting from a redistribution of confidence across the generation sequence.

4.2. Attribution Shift Towards Chain-of-Thought (Integrated Gradients)

Examination of attribution demonstrates that fine-tuned models functionally shift their attribution mass toward intermediate reasoning steps, validating their mechanistic reliance on the generated Chain-of-Thought. Figure 4 illustrates the distribution of attribution mass, measured via the length-normalized

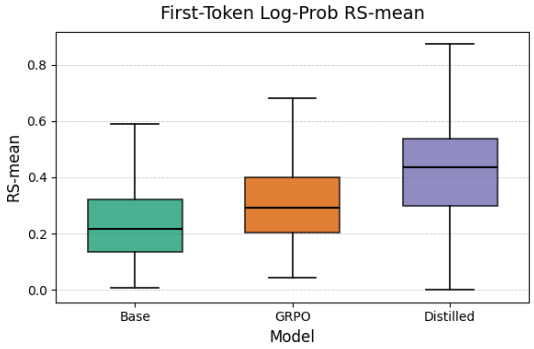


Figure 4: Functional importance of reasoning tokens for the first answer token prediction. The boxplots illustrate the distribution of relative attribution (RS-mean) across samples, where the central line denotes the median, the box spans the interquartile range (IQR), and whiskers extend to 1.5 × IQR. RL-aligned and Distilled models show a statistically significant increase in attribution mass toward the intermediate reasoning chain compared to the Base model, confirming a functional reliance on generated context for decision-making.

RS-mean metric to mitigate length bias from variable-length CoT generations. These attributions were calculated across the entire dataset, regardless of the final answer's correctness. When predicting the log-probability of the first answer token, both GRPO and Distilled models attribute substantially more importance to the intermediate reasoning steps than the Base model. The Distilled model shows the most aggressive shift (Base vs. Distilled: $p < 0.001$, Cliff's $\delta = -0.602$), effectively demonstrating a dependency on the preceding reasoning chain. GRPO also exhibits this functional shift, but with a slightly more distributed focus ($p < 0.001$, $\delta = -0.298$).

5. Discussion and Conclusion

In this paper, we investigated the opaque internal mechanisms behind the performance improvements of Reinforcement Learning Fine-Tuning (RLFT) in Large Language Models (LLMs). We proposed a mechanistic evaluation framework that leverages layer-wise dynamic extraction via the Tuned Lens and token-level attribution via Integrated Gradients (IG). By systematically comparing a base Qwen-2.5-14B model against its GRPO-fine-tuned and DeepSeek-R1 Distilled counterparts on the GSM8K dataset, this study provides a granular view of how Reinforcement Learning (RL) reshapes the internal computation and decision-making pathways of reasoning models.

A common assumption in current explainable AI literature is that RLFT inherently improves the core representational probability of the correct answer. However, our layer-wise analysis provides evidence to the contrary. The Tuned Lens evaluations (Fig. 3) show that the log-probability trajectory of the correct answer token remains nearly identical across all three model variants. Instead, the 7% accuracy improvement in the GRPO model stems from a delayed commitment. Specifically, the model maintains a high level of uncertainty throughout the reasoning phase, as evidenced by elevated intermediate entropy and forward KL divergence, before sharply converging at the output layers. Mechanistically, these findings suggest that rather than injecting novel knowledge into the latent space, RL operates as a mechanism that increases decisiveness at the output layer while tolerating internal uncertainty during reasoning.

This procedural delay is not the only internal change. Token-level attribution analysis shows that fine-tuned models also functionally shift their reliance toward their own intermediate reasoning steps. As illustrated in Fig. 4, while the base model relies mostly on the initial prompt, the reasoning models shift their attribution mass significantly toward the intermediate Chain-of-Thought (CoT) tokens. While expected for the Distilled model given its supervised training on 800,000 CoT traces, this reliance is particularly notable in the GRPO variant. Despite optimization exclusively via final-answer rewards without CoT examples exposure, the GRPO model shifted its internal attribution toward its self-generated reasoning chain. This indicates that functional CoT reliance is not merely a syntactic artifact but an

emergent mechanism driven by RL optimization.

Our findings should be interpreted in light of several constraints. While our evaluations focus exclusively on the arithmetic word problems in GSM8K, RL-enhanced reasoning improves performance on MATH dataset, MMLU-STEM technical knowledge, HumanEval coding, and even writing datasets. Further adaptation of our proposal could verify which conclusions hold for these distinct contexts. While literature indicates that RL-tuning frameworks yield performance gains across diverse and larger architectures, hardware constraints confined our current analysis to 14B-parameter models. These computational limits also necessitated reducing IG approximation to 15 steps for the longest Distilled sequences; while this could introduce minor attribution noise, the observed relative shifts remain statistically significant. Future work is required to verify if these internal procedural shifts scale linearly to larger models (such as 78B parameters) or manifest under alternative RL algorithms such as PPO. Lastly, because our Tuned Lens was trained on the general-domain Pile dataset, applying it to highly specialized mathematical generation may introduce representational distribution shifts.

Future research should integrate these findings into model training by utilizing mechanistic metrics, such as layer-wise forward KL divergence, as dense reward signals. Rewarding desirable internal computational paths, rather than solely evaluating final textual outputs, can yield safer and structurally robust reasoning systems.

Acknowledgments

The authors gratefully acknowledge the Instituto Tecnológico de Aragón (ITA) for providing high-performance-computing resources and expert technical support. This research was funded by the Spanish Ministry of Science and Innovation through the 2022 Public–Private Collaboration Programme, grant CPP2022-009790 (MAGELLAN: Design and deployment of a SaaS behavioural engine based on the ETR® AGI architecture), and by the Government of Aragón (Department of Science, University and Knowledge Society) under the Applied Research Groups scheme T17_23R (Research Group IODIDE: Energy-efficient systems and processes based on Artificial Intelligence and Power Electronics).

This work has also been partially funded by the Department of Big Data and Cognitive Systems at the Technological Institute of Aragon, under Retech Tourism-Spain Living Lab Agreement within the "PLAN FOR RECOVERY, TRANSFORMATION AND RESILIENCE - FINANCED BY THE EUROPEAN UNION - NEXTGENERATIONEU" and by the Government of Aragon.

All product names, trademarks and registered trademarks are property of their respective owners; their use herein does not imply endorsement.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3 in order to: Grammar and spelling check, as well as peer review simulation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[2] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[3] E. Zelikman, Y. Wu, J. Mu, N. D. Goodman, Star: self-taught reasoner bootstrapping reasoning with reasoning, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[4] E. Zelikman, G. R. Harik, Y. Shao, V. Jayasiri, N. Haber, N. Goodman, Quiet-STar: Language models can teach themselves to think before speaking, in: First Conference on Language Modeling, 2024.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[6] Y. Yue, Z. Chen, R. Lu, A. Zhao, Z. Wang, Y. Yue, S. Song, G. Huang, Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model?, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.

[7] M. Turpin, J. Michael, E. Perez, S. R. Bowman, Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[8] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for large language models: A survey, ACM Trans. Intell. Syst. Technol. 15 (2024). doi:10.1145/3639372.

[9] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, J. Schulman, Training verifiers to solve math word problems, 2021. doi:10.48550/arXiv.2110.14168.

[10] H. Zhang, Q. Hao, F. Xu, Y. Li, Reinforcement learning fine-tuning enhances activation intensity and diversity in the internal circuitry of LLMs, in: The Fourteenth International Conference on Learning Representations, 2026.

[11] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 3319–3328.

[12] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777. doi:10.5555/3295222.3295230.

[13] S. Abnar, W. Zuidema, Quantifying attention flow in transformers, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4190–4197. doi:10.18653/v1/2020.acl-main.385.

[14] L. Wang, L. Li, D. Dai, D. Chen, H. Zhou, F. Meng, J. Zhou, X. Sun, Label words are anchors: An information flow perspective for understanding in-context learning, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 9840–9855. doi:10.18653/v1/2023.emnlp-main.609.

[15] nostalgebraist, Interpreting GPT: The logit lens, LessWrong, 2020.

[16] N. Belrose, I. Ostrovsky, L. McKinney, Z. Furman, L. Smith, D. Halawi, S. Biderman, J. Steinhardt, Eliciting latent predictions from transformers with the tuned lens, 2025. doi:10.48550/arXiv.2303.08112. arXiv:2303.08112.

[17] Q. Zhang, D. Wang, H. Qian, Y. Li, T. Zhang, M. Huang, K. Xu, H. Li, L. Yan, H. Qiu, Understanding the dark side of LLMs' intrinsic self-correction, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 27066–27101. doi:10.18653/v1/2025.acl-long.1314.