**Attention Needs Intelligence: A Human Cognitive Inspired Approach to Enhance the Reasoning Capabilities of Large Language Models**

Vimukthi Senevirathna\*, Prashan Rathnayaka

School of Computing, Informatics Institute of Technology, 435 Galle Rd, Colombo 03, Sri Lanka

**ABSTRACT**

Research on reasoning capabilities of Large Language Models (LLMs) has been significantly improved throughout the past couple of years with the need of achieving human level intelligence in LLMs. Reasoning is the core skill that humans have which makes humans superior to other living beings in the activities of problem solving, critical thinking and decision making. Chain-of-Thoughts (CoT), Reinforcement Learning (RL) in LLM training pipeline and agent scratchpads were some of the approaches taken to pull out the reasoning capabilities of LLMs. Even though these methodologies show promising results, following a human cognitive inspired mechanism for language model reasoning can be seen as a significant gap in the domain, and yet the supervised fine-tuned behavior of these methodologies acts as a barrier for exploration and stochastic behavior of language models. A novel RL framework which aligns human cognitive principles and probabilistic reasoning to the LLM training pipeline will be introduced in this research study with the aim of achieving better reasoning capabilities through LLMs which have small number of parameters. The proposed solution is supplemented by a philosophically inspired novel CoT architecture and a Bayesian rewarding mechanism which gradually leads the LLM to follow probabilistic reasoning behavior. The trained policy model or the base LLM was successfully able to adhere to the introduced CoT structure by maintaining 89.5% accuracy for the gsmk8k and dolphin-r1 mixed test dataset. The reward distribution over training steps evidenced the expected stochastic reasoning pattern of the policy model without reward overoptimization.

**Keywords:** Reasoning in Large Language Models, Reinforcement Learning, Thompson Sampling, Chain-of-Thoughts

1. **INTRODUCTION**

Reasoning is the cognitive workflow which allows humans to solve problems, draw conclusions and make sense of information. This is fundamental across various domains such as science, mathematics, statistics and everyday life, to make informed decisions, solve complex problems, understand causal relationships with patterns and communicate effectively. Apart from drawing conclusions, reasoning enables humans to infer new knowledge from existing information and evaluate the validity of arguments. Reasoning can be categorized into two primary types, deductive reasoning and inductive reasoning. A specific conclusion made from general premises can be named as deductive reasoning. A popular example for this is, “since all humans are mortal, and Socrates is a human, Socrates is mortal”. Inductive reasoning on the other hand describes the pattern generalization from specific observations. As an example “The sun has risen every day to this day, therefore it will rise tomorrow as well” 1.

A key question in reasoning is whether it’s logical or probabilistic. Although logic is essential in reasoning, Philosophers have proven that it must be supplemented with probabilistic thinking to fully grasp the complexity of human cognition 2. The study argues for the concept of probabilistic validity, where the conclusions are validated based on their likelihood, rather than the logical necessity. Throughout the coming chapters this paper will be examining how these philosophical aspects of human cognition are incorporated into the LLMs training pipeline.

With the rise of LLMs such as Generative Pre-trained Transformer (GPT), Mistral and DeepSeek3, significant advancements have been made in enhancing reasoning capabilities through improved architectures. Most LLMs attempt to perform generalized reasoning with the patterns learned from a massive text corpus during training. In other words, it’s being called implicit reasoning where it often fails in complex multi-step reasoning problems. Explicit reasoning on the other hand facilitates CoT where a prompt guides the LLM to reason in a stepwise manner 4. Further, Reinforcement Learning from Human Feedback (RLHF) came into play where the models learn with human preferences, so that it gets chance to learn deterministic reasoning patterns in preference data 5.

While various reasoning techniques have been explored by the LLMs throughout the past couple of years, they often fail to learn generalized reasoning patterns in a probabilistic manner. Most existing methods rely on rigid rule-based or deterministic reasoning where it gets limited to diverse reasoning tasks. By taking all the facts into consideration, the proposed methodology will accommodate stochastic reason capabilities with diverse exploration into a language model which has 1.5 billion parameters with a novel CoT structure and a rewarding mechanism.

1. **RELATED WORK**

With the advancement of language models and promising results of Attention Mechanism, researchers have tried to achieve human level cognitive workflows through LLMs. Getting inspired by human cognition itself can be seen as one of the best avenues to achieve this. As discussed in the previous chapter, human reasoning weights more to the probabilistic reasoning, which can be achieved by applying Bayesian Networks in the algorithmic level of AI models 6. A recent work introduced a Bayesian reward model which signals higher uncertainties of reward score predictions7. The authors have trained a reward model using Low-Rank Adaptation (LoRA) weights and in the inferencing time they calculate the uncertainty of reward prediction by modeling the LoRA weights into a Posterior Distribution. Even though Bayesian modeling can be seen throughout the work, it has not been utilized for reasoning at an expected level.

Biological agents such as humans and animals exhibit advanced reasoning capabilities through hierarchical cognitive mechanism. This includes breaking down a complex problem into simple sub-tasks, gaining knowledge from prior experience and planning. Supplemented by this a recent study has proposed enhancements to be done in hierarchical reinforcement learning such as forward inverse models and hierarchical mental simulations 8. To overcome reasoning limitations in language models such as abstract planning where humans can plan using abstract representations and sequences, and composing known concepts in novel ways, researchers suggest that the attention mechanism should be inspired from the mechanism of Prefrontal Cortex (PFC) 9. This includes the modulation of attention mechanism, which should be more flexible and goal-conditioned, recurrence, and gating and serial processing of PFC, where these can be mimicked with Recurrent Neural Networks (RNNs) and Long Short-Term memory (LSTM), and dopamine guided RL.

Researchers shown that LLMs achieve emergent reasoning abilities by scaling up the size of model parameters, which cannot be seen in small language models 4. Apart from achieving reasoning capabilities with higher number of parameters, several novel frameworks, statistical methods and novel architectures have been introduced recently, where reasoning is promising with small language models as well. Concept of scratchpads has been introduced in LLMs to handle the tasks that require multistep computations 10. Scratchpad is sort of a buffer or memory where intermediate steps are generated and stored before producing the final answer. Instead of producing the final answer directly, the LLM generates a sequence of intermediate computation steps in the scratchpad, which allows the LLM to reason and perform tasks step-by-step. CoT is another approach taken to enhance the reasoning capabilities of LLMs where it provides intermediate steps for few-shot learning instead of simply providing input-output pairs 11. Abstracting from CoT a recent study showed that reasoning of LLMs can be drained out by simply adding the phrase ‘Let’s think step by step’ where it enables multi step reasoning process 12. Moving in advance, a methodology called Rethinking with Retrieval (RR) has been introduced where it generates multiple diverse reasoning paths for a given query 13. Then these steps are used to guide knowledge retrieval from external sources where it acts as search queries. Finally, the model selects the answer supported by the most reliable reasoning path. Yet, following a human cognitive inspired reasoning structure remains unaddressed.

1. **METHODOLOGY**

This work offers a novel RL training pipeline which introduces the stochasticity to the reasoning of the policy model via reward assignment while maintaining a philosophically inspired CoT structure. The proposed methodology avoids policy model by learning rigid reasoning patterns by adapting stochastic reasoning behavior, in contrast to current RL methods which learn or overoptimize deterministic reasoning.

**3.1 Novel CoT Reasoning Dataset**

The proposed CoT architecture was inspired by an early stage work, where authors have identified and proposed a cognitive architecture in human reasoning and planning 14. The authors have identified four key components involving human thinking which can be used in computer science via novel data architecture. Structured reasoning is important in LLM inferencing to boost up the exploration with a forceful step wise token generation with already learned patterns.

A synthetic dataset has been generated which aligns with the novel key components, from the original dataset ‘cognitivecomputations/dolphin-r1’ which is available under Hugging Face. GPT-4o-mini has been used for the synthetic dataset generation process with an instructional prompt. The dataset breakdowns the answer into four step wise reasoning steps, for a given question. Goal detector is the initial step where LLM identifies its purpose properly, then plan generator comes into play where LLM generates a plan to achieve the identified goal. Projector, where LLM projects whether the generated plan leads to a successful conclusion. Finally, the executor, where LLM executes the plan and concludes if the projector is confident about the plan.

*Ex: - Question: What is the sum of 2 + 2?*

*Answer: <goal\_detector> Identify the problem goal </goal\_detector>*

*<plan\_generator> Generate a stepwise solution plan </plan\_generator>*

*<projector> Predict the outcome of executing the plan </projector>*

*<executer> Generate and conclude the final answer </executer>*

**3.2 Introducing Probabilistic Reasoning with Thompson Sampling**

Since the proposed solution is fully inspired by the philosophical aspect of human cognition, introducing probabilistic reasoning has been identified as a key component when aligning LLMs to human cognition. Achieving probabilistic reasoning with a base LLM architecture change is computationally expensive and time consuming. In that case a framework has been introduced to lead the base LLM for probabilistic reasoning by the influence of reward function during the training.

**3.2.1 Training the Reward Model**

The “distilroberta-base” which is available under Hugging Face has been used to train the reward model. This 70 million parameterized model allows efficient training and pattern adaptation in about two hours in a T4 GPU with the synthetically generated dataset. A portion from the synthetic dataset has been used to train the reward model with 9:1 ratio of train and validation split. The dataset includes three columns, instruction, chosen response and rejected response, where chosen response includes the novel CoT structured answer and rejected response includes the original answer of the dataset. Hugging Face ‘RewardTrainer’ class has been used to train the reward model where it learns to give higher positive floating values to the answers that follows novel CoT structure and negative values to the plain reasoning templates. By aligning with the CoT tags and penalizing the unsupported logical jumps, the reward model’s weightings, which prioritize structure over content, are implicitly learned.

**3.2.2 Applying Thompson Sampling for Probabilistic Reasoning**

As discussed at the beginning of this chapter, this work incorporates probabilistic reasoning through reward assignment by introducing a computationally efficient way to align probabilistic human reasoning behavior. Existing reward functions assign the rewards as they are, during the RL pipeline where it leads to deterministic behavior of the model through over optimized weights. First, the reward model has been defined as a Bayesian model, which allows to calculate uncertainty in reward scores by modeling the reward scores as a probabilistic distribution. The posterior distribution over the reward weights (ΔW) is given by,

Here, is the likelihood of data given that reward weights , which describes how well the weights explain the data, is the prior distribution over the weights and is the posterior, representing belief over the reward function’s parameters after observing data.

During training, a finding has been arisen where modeling the posterior directly is computationally expensive with the calculations of second-order derivatives of the loss function with respect to the model parameters. In that case, the posterior was approximated using the Fisher Information, which allows for efficient uncertainty estimation. Thus, instead of explicitly computing uncertainty with , the uncertainty has been approximated in the reward model using as follows.

Where the gradient of the reward model for weight, is represented by , and is the Fisher Information Matrix diagonal value for weight. Fisher Information Matrix is able to provide an approximation of the model’s sensitivity to small changes in parameters which indicates how confident the model is about its learned parameters.

With this approach, the likelihood is still implicitly included via the gradient-based Fisher Information and yet the prior is often assumed to be non-informative (flat prior) in Laplace Approximation.

Finally, to polish the reward assignment with stochasticity with the aim of introducing probabilistic reasoning behavior in policy model, the Tompson Sampling was applied to the calculated uncertainty. The final Thompson Sampled reward is defined by ensuring probabilistic exploration as below.

Where R is the raw reward score obtained from the reward model.

A diagram of a computer

AI-generated content may be incorrect.

Figure 1: High-level architecture diagram of proposed reward sampling mechanism

The sampled reward will be returned as the finalized reward during the training of policy model. With this novel approach the proposed solution was able to manipulate the reasoning procedure of the policy model with the Thompson Sampled dynamic reward assignment.

**3.2.3 Training Policy Model with Group Relative Policy Optimization (GRPO)**

Qwen2.5-1.5B-Instruct, which is a 500 million parameterized model, has been used as the policy model to apply training with proposed framework. Thousand questions from the same dataset have been taken to train the base model with the novel framework. ‘GRPOTrainer’ class from Hugging Face has been used during training which incorporates Group Relative Policy Optimization (GRPO) as the principal RL algorithm. Training to a one epoch in a single A100 GPU with LoRA weights, a learning rate of 0.0001, and two generations per query introduced expected reasoning results which will be demonstrated in the coming chapter.

1. **EXPERIMENTS AND RESULTS**

**4.1 Reward Model Evaluation**

Figure 1 illustrates the progression of training loss and validation loss during the fine-tuning process of reward model. The loss starts at 0.7 and smoothly decreases to near 0.05, indicating effective leaning and the closely aligned curves suggest that the model isn’t memorizing the data but generalizing well to unseen samples.

**A graph with blue and orange lines

AI-generated content may be incorrect.**

Figure 2: Training and Validation loss of Reward Model

**4.2 Policy Model Evaluation**

Figure 2 shows the KL divergence throughout the training steps. A sharp increment can be seen initially, indicating that the model is rapidly adapting its policy distribution in response to the learned reward function. Moving forward with the training divergence stabilizes suggesting that the policy model is aligning well with the reward distribution while maintaining a controlled level of exploration. On the other hand, the reward steadily increases as training progresses, demonstrating that the policy model is successfully optimizing towards responses that align with the expected reasoning format while maintaining a stochastic reasoning behavior.

A graph of a graph

AI-generated content may be incorrect.

Figure 3: KL divergence, reward distribution and training loss of the policy model

The fine-tuned policy model has been compared and validated over the gsm8k and dolphin-r1 mixed test set where it depicts the same accuracy as the base instruct model, suggesting that the introduced stochastic reasoning behavior achieved the expected reasoning capabilities with maintaining the accuracy as it is. Further, LLMs-evaluating-LLMs technique has been used to efficiently evaluate the stochastic reasoning behavior of the fine-tuned model and the base language model. GPT-4o model has been used with an instructional prompt to evaluate the stochastic reasoning behavior of both the models by passing 20 batches, each batch containing 10 questions. After averaging the fine-tuned model got a score of 7.9 for the stochastic reasoning patterns and the base language model got a score of 5.6 which concludes that the proposed solution was successfully able to follow a stochastic reasoning pattern.

1. **CONCLUSION**

In this work, the core objective was to enhance the reasoning capabilities of small-parameterized language models throughout a human cognitive inspired training framework. A novel CoT structure has been introduced to mimic the reasoning steps of a human, and a Thompson sampling-based reward mechanism was introduced to guide the policy model for stochastic reasoning behavior. Key findings of the work include a well-trained reward model and a stochastic policy model with the expected KL-divergence and reward distribution. Although the policy model which was fine-tuned upon the proposed framework has the same level of accuracy to its base model, an expected level of stochastic reasoning behavior was observed throughout the fine-tuned policy model. Further, the ethical implications must be considered, particularly how unbalanced dataset coverage or reward bias may interrupt the probabilistic reasoning process. Ensuring transparency in reward model is a must when deploying such systems in sensitive domains. Focusing on improving accuracy while maintaining the same stochastic reasoning structure with more computational resources can be proposed as possible future enhancement.

**REFERENCES**

[1] Huang, J., and Chang, K.C.-C., “Towards Reasoning in Large Language Models: A Survey,” arXiv:2212.10403, arXiv (2023).

[2] Johnson-Laird, P.N., Khemlani, S.S., and Goodwin, G.P., “Logic, probability, and human reasoning,” Trends in Cognitive Sciences 19(4), 201–214 (2015).

[3] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., et al., “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning,” arXiv:2501.12948, arXiv (2025).

[4] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., et al., “Emergent Abilities of Large Language Models,” arXiv:2206.07682, arXiv (2022).

[5] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D., “Deep Reinforcement Learning from Human Preferences.”

[6] Oaksford, M., and Chater, N., “The probabilistic approach to human reasoning,” Trends in Cognitive Sciences 5(8), 349–357 (2001).

[7] Yang, A.X., Robeyns, M., Coste, T., Shi, Z., Wang, J., Bou-Ammar, H., and Aitchison, L., “Bayesian Reward Models for LLM Alignment,” arXiv:2402.13210, arXiv (2024).

[8] Eppe, M., Gumbsch, C., Kerzel, M., Nguyen, P.D.H., Butz, M.V., and Wermter, S., “Intelligent problem-solving as integrated hierarchical reinforcement learning,” Nature Machine Intelligence 4(1), 11–20 (2022).

[9] Russin, J., O’Reilly, R.C., and Bengio, Y., “DEEP LEARNING NEEDS A PREFRONTAL CORTEX” (2020).

[10] Nye, M., Andreassen, A.J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., et al., “SHOW YOUR WORK: SCRATCHPADS FOR INTERMEDI- ATE COMPUTATION WITH LANGUAGE MODELS.”

[11] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., and Zhou, D., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.”

[12] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., and Iwasawa, Y., “Large Language Models are Zero-Shot Reasoners.”

[13] He, H., Zhang, H., and Roth, D., “Rethinking with Retrieval: Faithful Large Language Model Inference,” arXiv:2301.00303, arXiv (2022).

[14] Wilensky, R., “Planning and understanding: A computational approach to human reasoning” 23(2), (1984).