Do LLMs Dream of Electric Emotions? Towards Quantifying Metacognition and Generalizing the Teacher-Student Model Using Ensembles of LLMs

Ricky J. Sethi
Fitchburg State University, Worcester Polytechnic Institute, National University
Fitchburg, MA, USA

Hefei Qiu
Fitchburg State University
Fitchburg, MA, USA

Charles Courchaine
Fitchburg State University, National University
Fitchburg, MA, USA

Josh Iacoboni
Fitchburg State University
Fitchburg, MA, USA

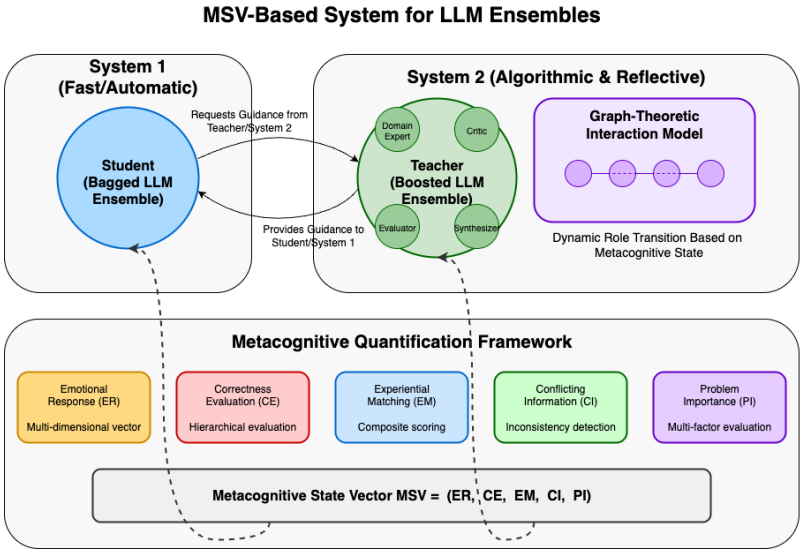

Figure 1: Metacognitive Framework for LLM Ensembles

Abstract

In this paper, we propose a novel framework for quantifying metacognitive processes in ensembles of Large Language Models (LLMs) and extending the traditional teacher-student model through the lens of dual-process cognitive theory. We introduce a Metacognitive State Vector (MSV) that operationalizes metacognition across five dimensions: emotional response analysis, correctness evaluation, experiential matching, conflicting information estimation, and problem importance task prioritization.

In our formulation, the rapid, intuitive thinking of System 1 is mapped onto a smaller "student" (single LLM or ensemble of bagged LLMs) while the deliberate, analytical reasoning of System 2 is mapped onto a larger "teacher" (ensemble of boosted LLMs) using the MSV. Additionally, we utilize a graph-theoretic architecture to model ensemble interactions, enabling LLMs to assume dynamic roles and transition between System 1 and System 2 for improved decision-making. We view this work as a first step towards establishing a true measure of emergent metacognition in such systems.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3760839

CCS Concepts

• **Computing methodologies** → *Cognitive science*; *Ensemble methods*; • **Theory of computation** → *Machine learning theory*.

Keywords

Large Language Models, Metacognition, LLM Ensemble Methods, LLM Emotions, Teacher-Student Model, Dual-Process Theory, Cognitive Psychology, Neuroscience, Neural Framework

ACM Reference Format:
Ricky J. Sethi, Hefei Qiu, Charles Courchaine, and Josh Iacoboni. 2025. Do LLMs Dream of Electric Emotions? Towards Quantifying Metacognition and Generalizing the Teacher-Student Model Using Ensembles of LLMs. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3760839

1 Introduction

1.1 Background and Motivation

The recent advent of Generative AI (GenAI) systems like Large Language Models (LLMs) has transformed how we solve problems [1, 10, 11] but our understanding of their decision-making processes and metacognitive capabilities is significantly limited [3, 14]. In addition, traditional approaches to model distillation and ensemble methods have primarily focused on performance metrics rather than the qualitative aspects of decision-making that characterize human cognition [6, 13].

In this paper, we propose to bridge this gap by introducing a framework, shown in Figure 1, that integrates the Dual-Process Cognitive Theory from neuroscience with LLM ensemble architectures [8] as a way to generalize the traditional Teacher-Student Model [6] as well as to quantify metacognition in LLMs. Our framework helps connect GenAI and human cognition by providing a theoretical foundation for quantitatively measuring metacognition in AI systems using a 5-dimensional Metacognitive State Vector (MSV). This MSV incorporates emotional responses, correctness evaluations, and problem importance assessments by mapping neuroscience principles to LLM ensembles, thereby advancing our understanding of how AI systems can better mimic human metacognitive abilities.

1.2 Current Approaches and Limitations

Knowledge distillation in LLMs is often used to transfer knowledge from a large, complex model (the teacher) to a smaller, more efficient model (the student). The teacher model is usually trained on a comprehensive dataset and then the student model is trained using a combination of the original training data and the soft targets provided by the teacher model; these soft targets are the probabilities assigned to each class by the teacher model, which contain more information than the hard targets [9, 15]. Current Teacher-Student models in machine learning primarily focus on knowledge transfer through probability distribution matching [6]. This kind of knowledge distillation allows the student model to learn from the nuanced patterns captured by the teacher model and can lead to improved generalization, allowing the student model to have significantly fewer parameters.

In fact, recent research has explored the use of neurobiologically-inspired neural networks to improve the performance of student models by leveraging insights from neuroscience to design architectures that mirror the cognitive processes of the human brain, with preliminary results in the context of math problem solving suggesting significant potential for enhancing the capabilities of student models [4, 9]. These models can also help transfer the metacognitive skills of a strong LLM to a weaker one, thus enhancing the student's performance on various tasks [4]. Metacognition, often called "thinking about thinking" in cognitive psychology, involves the ability to evaluate and regulate one's own cognitive processes. In the context of LLMs, this can be quantified through various methods such as emotional response analysis, correctness evaluation, experiential matching, and problem importance assessment, where these methods provide insights into how LLMs can mimic human metacognitive abilities, thus enhancing their decision-making capabilities [1, 13].

1.3 Research Objectives

Our aim is to extend beyond these current approaches in order to provide better understanding and deeper information processing and knowledge representation of decision-making processes using ensembles of LLMs. As such, we propose a computational framework to enhance the sophistication and transparency of decision-making in LLM ensembles and more closely approximate human cognitive capabilities.

Our proposed framework has three primary objectives:

(1) Establish **quantifiable metrics for metacognitive processes** in LLMs to help *operationalize cognitive assessment* as a Metacognitive State Vector (MSV), as seen in Section 2.2
(2) Map the System 1 and System 2 cognitive processes onto distinct LLM ensemble architectures using the MSV to govern role engagement and **extend the teacher–student paradigm** into a dual-process framework for *collaborative reasoning* at inference-time as well as training-time, as seen in Section 2.1
(3) Introduce a graph-theoretic control model for LLM ensembles that allows for **dynamic role transition** mechanisms which are grounded in *cognitive neuroscience principles* for System 1/2 switching, as seen in Section 2.3

We see the Teacher-Student model as a special case of this generalized Dual-Process Cognitive framework consisting of System 1 (Student) and System 2 (Teacher) that allows for a more nuanced understanding of knowledge transfer and metacognitive capabilities of LLMs. We adopt the System 1/System 2 framework as an explanatory scaffolding, where System 1 denotes a fast, low-cost path that yields quick answers and System 2 denotes a slower, higher-cost refinement process. Concretely, our implementation realizes these as *parallel aggregation* and *sequential refinement* compute regimes selected by control signals as determined by the MSV.

In addition, the MSV aligns with classic software engineering principles, where its aggregate value can potentially serve as a surrogate *validation* metric for the fit of the response to the initial problem, while its individual components can provide granular metrics for *verification* of specific aspects of the response, thus bridging cognitive science and established software QA frameworks.

To the best of our knowledge, this framework represents a significant methodological advancement in conceptualizing artificial metacognition, providing both theoretical foundations and practical implementation strategies. Importantly, we do not claim that this work demonstrates emergent metacognition in the human sense of "thinking about thinking." Rather, we present it as a control framework for resource allocation and answer improvement and a potential stepping stone towards a measurable version thereof.

2 Theoretical Framework

2.1 Dual-Process Cognitive Model

Our framework builds upon Stanovich's tri-process theory, which distinguishes between System 1 (the Autonomous Set of Systems (TASS)) and System 2 (the algorithmic mind and the reflective mind) [13]. We map these components onto an LLM ensemble architecture for each where System 1 is implemented as a bagged LLM ensemble to smooth out predictions and System 2 is implemented as a boosted LLM ensemble with defined roles and evaluation metrics for deeper understanding and reasoning, as shown in Figure 1.

System 1 is characterized by fast, automatic, and often subconscious processes while System 2 is slower, more deliberate, and conscious. System 1 processes tend to be quick and rely on heuristics, making them efficient for routine tasks and immediate responses; System 2 processes, on the other hand, involve analytical thinking, requiring more cognitive resources and time, and are engaged when tasks demand careful consideration, logical reasoning, and the evaluation of complex problems [13].

In the context of LLMs, mapping these cognitive systems onto AI architectures involves distinguishing between tasks that can be handled by pre-trained, automatic responses (analogous to System 1) versus those that require more sophisticated, context-aware processing (analogous to System 2). For instance, an LLM might use System 1-like processes to generate quick responses based on pattern recognition and past data, while System 2-like processes might be invoked for tasks requiring deeper understanding and reasoning, such as evaluating the correctness of a response or assessing the importance of a problem [10].

We model the interaction between these systems dynamically using graph-theoretic approaches where we represent the ensemble of LLMs as nodes in a graph, with edges denoting the interactions and information flow between them and each node dynamically taking on a role determined by the weights of its **Metacognitive State Vector (MSV)**, as shown in Section 2.3; these **roles** can have labels like Domain Expert, Critic, Evaluator, Synthesizer, etc. This allows us to model how System 1 and System 2 processes influence each other, providing insights into the emergent behavior of the ensemble. For example, a graph-theoretic model can help visualize how a quick, heuristic-based decision (System 1) might trigger a more detailed, analytical review (System 2) when certain conditions are met, such as the detection of a potential error or the need for a more nuanced response [1, 2, 5].

The MSV has five dimensions which quantify metacognitive processes in LLMs on a common [0,100] scale, although implementations may optionally normalize to [0,1] prior to activation (e.g., sigmoid/softmax). These metrics are essential for understanding how LLMs can mimic human-like metacognition, where System 1 might generate an initial emotional response or heuristic judgment, and System 2 might subsequently evaluate and refine this response based on additional information and logical analysis [10, 13]. In addition, by modeling these complex and multi-faceted interactions dynamically and quantifying metacognitive processes, we can enhance the decision-making capabilities of AI systems, making them more robust and sophisticated in handling a wide range of tasks. This approach not only advances our understanding of AI but also provides a deeper insight into the cognitive processes that underlie human intelligence [2, 7].

2.2 Metacognitive State Vector (MSV)

We propose a comprehensive framework for quantifying metacognitive processes across five primary dimensions, each implemented through specific computational mechanisms as follows.

2.2.1 Emotional Response (ER) Implementation. The ER metric captures the affective components of cognitive processing by aggregating the emotion intensities $\{v_1, v_2, \ldots, v_n\}$ for n distinct emotions, each bounded in $[0, 100]$ [12], and the overall response is:

$$ER = \sum_{i=1}^{n} \epsilon_i v_i, \quad \text{with } \epsilon_i \geq 0, \ \sum_{i=1}^{n} \epsilon_i = 1, \tag{1}$$

with context-dependent weights ϵ_i. ER is pivotal in prioritizing information and identifying potential ethical concerns; high ER values typically accelerate System 2 engagement. This can be implemented via sentiment analysis using fine-tuned emotion recognition models, contextual valence shifters detection, physiological response emulation through token-level analysis, ensemble agreement, etc.

2.2.2 Correctness Evaluation (CE) Framework. CE implements a hierarchical evaluation structure:

$$\begin{aligned} CE = &\alpha_1 F_1(\text{logical\_consistency}) + \alpha_2 F_2(\text{factual\_accuracy}) + \\ &\alpha_3 F_3(\text{contextual\_appropriateness}) \end{aligned} \tag{2}$$

where F_1, F_2, F_3 are evaluation functions mapping to $[0, 100]$, $\alpha_1, \alpha_2, \alpha_3$ are context-dependent weights, and $\sum_i \alpha_i = 1$. The CE dimension provides a comprehensive assessment of response quality confidence across multiple validity dimensions and aligns with human epistemic monitoring. We convert this confidence score signal to an $uncertainty\_signal = 1 - CE$ where low values of the uncertainty signal serve as robust triggers for System 2 intervention, initiating more deliberative processing and potential role transitions to domain experts or critics within the ensemble.

2.2.3 Experiential Matching (EM) System. EM quantifies alignment through:

$$\begin{aligned} EM = &\omega_1 K(\text{response}, \text{knowledge\_base}) + \\ &\omega_2 S(\text{response}, \text{historical\_responses}) \end{aligned} \tag{3}$$

where $K(\cdot, \cdot)$ measures knowledge base similarity, $S(\cdot, \cdot)$ computes historical response similarity, and ω_1, ω_2 are adaptive weights based on context. The EM dimension models relating new information to existing/familiar knowledge structures, drawing parallels to episodic and semantic memory integration in human cognition, leveraging past experiences. We convert this familiarity score signal to an $unfamiliarity\_signal = 1 - EM$ where low values of the unfamiliarity signal can trigger exploration of alternative knowledge sources, while high scores reinforce confidence in generated responses.

2.2.4 Conflicting Information (CI) Estimation. CI quantifies the degree of inconsistency and contradictory information present:

$$\begin{aligned} CI = &\delta_1 D(\text{internal\_consistency}) + \delta_2 D(\text{source\_agreement}) + \\ &\delta_3 D(\text{temporal\_stability}) \end{aligned} \tag{4}$$

where D(internal_consistency) measures logical contradictions within a single response, D(source_agreement) evaluates disagreement across multiple sources, D(temporal_stability) assesses consistency of information over time, and δ_i are context-sensitive weighting coefficients with $\sum \delta_i = 1$. The CI dimension serves as a critical trigger for System 2 activation, as high conflict signals the need for deeper analytical processing and aligns with cognitive science research showing that humans engage deliberative reasoning when confronted with contradictory information.

2.2.5 Problem Importance (PI) Task Prioritization Assessment. PI implements a multi-factor evaluation:

$$\begin{aligned} PI = &\beta_1 C(\text{potential\_consequences}) + \\ &\beta_2 U(\text{temporal\_urgency}) + \beta_3 I(\text{scope\_impact}) \end{aligned} \tag{5}$$

where β_i are the dynamically adjusted weights with $\beta_k >= 0$ and $\sum \beta_k = 1$, $C(\cdot)$ evaluates potential consequences, $U(\cdot)$ measures temporal urgency, and $I(\cdot)$ assesses scope of impact. The PI dimension captures the critical metacognitive function of resource allocation based on task significance and parallels the human ability to prioritize cognitive resources toward high-stakes or time-sensitive situations; high PI scores trigger more extensive System 2 processing, activating additional ensemble members and enabling more thorough verification procedures.

2.3 Role Transition Dynamics

The interaction between ensemble members is modeled using the **graph-theoretic ensemble interactions** graph, which is defined as $G(V, E, W)$ where $V = \{v_1, \ldots, v_n\}$ represents the set of LLM nodes and $E \subseteq V \times V$ is the subset of all possible pairs of nodes where each element $e \in E$ is a directed edge from one node to another. In addition, each edge is assigned a weight $W(e) = (w(e), \mu(e)(M))$ depending on $w(e)$, the base weight of edge e that is independent of the metacognitive state, as well as the metacognitive transition function, $\mu(e)$, for edge e that depends upon its **Metacognitive State Vector (MSV)**, $M = (ER, CE, EM, CI, PI)$.

Each node v_i updates its role based on the weight information it receives with probability, $P(r'|r, M) = softmax(T(r, r', M))$, where r is the current role and r' is the new role. Thus, the weights on the MSV determine the role(s) each node assumes.

Finally, the **Formal Role Transition Framework** $\Gamma = (R, S, T, M)$ represents the complete role transition system where $R = \{r_1, \ldots, r_k\}$ is the set of possible roles, $S = \{s_1, \ldots, s_n\}$ is the set of system states, and $T : R \times S \to \mathcal{P}(R)$ defines allowable transitions. The metacognitive dimensions work in conjunction with each other to determine when System 2 processes should be activated, which roles should be engaged, and how ensemble resources should be allocated to optimize response quality.

Analogy: A helpful analogy for this system would be a research team (ensemble of LLMs) where each person (LLM node) has their own mental (metacognitive) state and confidence levels. In addition, each person decides their role based on their own mental (metacognitive) state and the team's overall approach (the ensemble role designation) depends on how the individual roles are combined. Information sharing between people (the nodes) affects their individual role decisions and the overall team (ensemble) designation.

Node	Role	Metacognitive State Vector				
		ER	CE	EM	PI	CI
Node_1	Critic	20	30	70	50	80
Node_2	Expert	10	90	85	70	20
Node_3	Critic	40	25	60	40	75

Table 1: Initial node configuration and role assignments.

Ensemble Dynamics for an Example System: Suppose we have three nodes with the initial metacognitive state vector configuration shown in Table 1. The graph structure is defined as:

$$V = \{\text{Node}_1, \text{Node}_2, \text{Node}_3\}$$
$$E = \{(\text{Node}_1, \text{Node}_2), (\text{Node}_2, \text{Node}_3), (\text{Node}_3, \text{Node}_1)\}$$
$$W(e) = (w(e), \mu(e)) \text{ for each } e \in E$$

The metacognitive activation function $\mu(e)(M_i)$ for edge e originating from node i is computed as:

$$\mu(e_{ij})(M_i) = \sigma(\alpha_1 \cdot ER_i + \alpha_2 \cdot CE_i + \alpha_3 \cdot EM_i + \alpha_4 \cdot PI_i + \alpha_5 \cdot CI_i)$$

where $\sigma(x) = \frac{1}{1+e^{-x/\tau}}$ is the sigmoid function with τ being the temperature parameter which controls the steepness of the sigmoid function. Since the metacognitive values are in [0,100] and weighted sums typically range 20-80, setting $\tau = 10$ maps this to [2,8], which gives a nice range of sigmoid outputs; alternatively, with normalization of all inputs, the standard sigmoid can be used with $\tau = 1$. Similarly, α are the edge-specific weights and might be set to something like $\boldsymbol{\alpha} = (0.1, 0.2, 0.1, 0.3, 0.3)$.

In regards to role transition analysis, suppose Node_1 (initially Critic) considers a transition to a new role, Expert. The transition scoring function would then be:

$$T(r, r', M_i) = w_1 \cdot ER_i + w_2 \cdot CE_i + w_3 \cdot EM_i + w_4 \cdot PI_i + w_5 \cdot CI_i$$

For the transition Critic $\to$ Expert, we might use role-specific weights like $\boldsymbol{w}_{\text{Critic}\to\text{Expert}} = (0.1, 0.4, 0.1, 0.2, 0.2)$, which are learned or set by domain experts. E.g., here higher weight is placed on Correctness Evaluation (CE), Conflict Information (CI), and Problem Importance (PI) as these dimensions indicate the need for expert consultation. The role transition probability is then computed using the softmax function:

$$P(\text{Expert}|\text{Critic}, M_1) = \frac{\exp(T(\text{Critic}, \text{Expert}, M_1))}{\sum_{r' \in \text{Roles}} \exp(T(\text{Critic}, r', M_1))}$$

The ensemble's overall role designation is determined through weighted aggregation:

$$\text{Ensemble Role} = \arg\max_R \sum_{i:\text{role}_i = R} \text{confidence}_i$$

where confidence_i is derived from the maximum CE value. Therefore, each node evaluates potential role transitions using received information and its current metacognitive state and the overall ensemble role emerges from the weighted combination of individual node roles and confidence levels. This framework thus enables dynamic adaptation where nodes with higher confidence (like Node_2 with CE=90) have greater influence on ensemble behavior, while metacognitive conflicts (like high CI or PI values) trigger appropriate role and System 1/System 2 transitions to address uncertainty and maintain system coherence.

3 GenAI Usage Disclosure

GenAI tools were not used for the writing, code, or data collection.

References

[1] Mehrdad Ashtiani and Mohammad Abdollahi Azgomi. 2015. A survey of quantum-like approaches to decision making and cognition. *Mathematical Social Sciences* 75 (3 2015). https://doi.org/10.1016/j.mathsocsci.2015.02.004

[2] Jelle Bruineberg, Julian Kiverstein, and Erik Rietveld. 2018. The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese* 195 (10 2018). https://doi.org/10.1007/s11229-016-1239-1

[3] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, and et al. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 15, 3 (2024), 39. https://doi.org/10.1145/3641289

[4] Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. 2024. Metacognitive Capabilities of LLMs: An Exploration in Mathematical Problem Solving. https://doi.org/10.48550/arXiv.2405.12205

[5] Emmanuel Haven and Andrei Khrennikov. 2016. Statistical and subjective interpretations of probability in quantum-like models of cognition and decision making. *Journal of Mathematical Psychology* 74 (3 2016). https://doi.org/10.1016/j.jmp.2016.02.005

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML] https://arxiv.org/abs/1503.02531

[7] Sven Hoffmann and Christian Beste. 2015. A perspective on neural and cognitive mechanisms of error commission. *Frontiers in Behavioral Neuroscience* 9, 50 (3 2015). https://doi.org/10.3389/fnbeh.2015.00050

[8] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025. From System 1 to System 2: A Survey of Reasoning Large Language Models. January (2025), 1–35. arXiv:2502.17419 http://arxiv.org/abs/2502.17419

[9] Fuseini Mumuni and Alhassan Mumuni. 2024. Improving deep learning with prior knowledge and cognitive models: A survey on enhancing explainability, adversarial robustness and zero-shot learning. *Cognitive Systems Research* (2024). https://doi.org/10.1016/j.cogsys.2023.101188

[10] B. Porter, V. Lifschitz, and F. van Harmelen (Eds.). 2007. *Handbook of Knowledge Representation*. Elsevier B.V.

[11] Ricky J Sethi. 2020. Essential Computational Thinking: Computer Science from Scratch.

[12] Ricky J. Sethi, Raghuram Rangaraju, and Bryce Shurts. 2019. Fact Checking Misinformation Using Recommendations from Emotional Pedagogical Agents. In *Intelligent Tutoring Systems*. Springer International Publishing.

[13] Keith E. Stanovich. 2008. Chapter 3 Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?. In *Proceedings of the Cognitive Science Conference*.

[14] Yuqing Wang and Yun Zhao. 2024. Metacognitive Prompting Improves Understanding in Large Language Models, In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. https://github.com/EternityYW/Metacognitive-Prompting

[15] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large Language Models as Optimizers, In ICLR. *arXiv:2309.03409v3 [cs.LG]*, 1–15. https://arxiv.org/abs/2309.03409v3

Received 06 June 2025; accepted 04 August 2025