Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084480

ARTICLE

Research on an Emergence Mechanism in Large Language Models for Command and Decision-Making

Yazhi Zheng[1,2], **Xiaolong Cui**[1,*], **Xin Wang**[1,2,#] **and Xuanzhu Sheng**[1,2,#]

[1]Key Laboratory of CTC & IE, Ministry of Education, Engineering University of PAP, Xi'an, China
[2]Graduate Brigade, Engineering University of PAP, Xi'an, China

*Corresponding Author: Xiaolong Cui. Email: xlspace@foxmail.com

[#]These authors contributed equally to this work

Received: 23 April 2026; Accepted: 03 August 2026; Published: 15 September 2026

ABSTRACT: Large Language Models (LLMs) currently lack the robust command and decision-making (C&D) capabilities essential for the command and control domain. To address this critical gap, this paper proposes an emergence mechanism that integrates a domain-specialized Chain of Thought (CoT) framework with a Process Reward Model (PRM)-inspired evaluation and inference-time optimization paradigm. We construct a novel Chain of Command and Decision (CoCD) framework, a C2-specific CoT structure with contextual persistence, knowledge accumulation, and a human-in-the-loop feedback loop, and define a four-dimensional PRM-inspired evaluation framework for process-level assessment of C&D reasoning. Experimental evaluations on 40 C&D scenarios of varying complexity demonstrate that the CoCD framework significantly outperforms direct prompting (Mann–Whitney $U = 1314$, $p < 0.0001$, Cohen's $d = 1.340$) and Standard-CoT ($p = 0.005$, $d = 0.606$) in composite performance. PRM-guided Best-of-N selection further improves performance by 5.8% over single-sample CoCD ($p < 0.001$, $d = 0.855$), providing direct empirical evidence for the utility of process-aware reward signals at inference time. CoCD's structural advantage is greatest in high-uncertainty, structurally ambiguous scenarios (Level 3 gap: +0.925 points), revealing a complexity-type effect that informs the deployment scope of structured CoT frameworks. These findings provide empirical support for domain-specialized structured reasoning and process-level evaluation as foundations for future RL-based C&D capability development in LLMs.

KEYWORDS: Large language models; chain of thought; process reward model; chain of command and decision; emergence mechanism; Best-of-N selection

1 Introduction

Command and decision-making (C&D) is the cognitive core of military command and control (C2) systems, requiring commanders to synthesise multi-source intelligence, balance competing objectives, and produce executable decisions under time pressure and uncertainty. The emergence of Large Language Models (LLMs) with strong complex reasoning capabilities creates an unprecedented opportunity to automate or augment this process. However, realising this potential is non-trivial: C2 tasks demand that reasoning steps be traceable, domain-aligned, and reliably structured—properties that generic LLM outputs lack, because the CoT generated by unconstrained LLMs often follows no stable logical order and provides no handle for quality verification.

Existing research addresses LLM reasoning through two main paths, both of which fall short for C&D. CoT prompt engineering [1,2] guides models toward step-by-step reasoning, but generic templates (e.g., "Let's think step by step") carry no military domain knowledge and cannot enforce the specific inferential structure that C2 logic requires. Reinforcement Learning (RL)-based optimisation [3] can refine model outputs through reward feedback, but process reward models (PRMs) that evaluate intermediate reasoning steps have only recently been explored [4–6], and no prior work has adapted them to the C&D domain. Critically, neither line addresses the joint need for domain-structured CoT and process-level evaluation that together can underpin measurable, controllable C&D capability improvement.

To close this gap, this paper proposes a CoCD-based emergence mechanism consisting of two implemented components. First, we design the Chain of Command and Decision (CoCD) framework, a four-node C2-specific CoT scaffold (Situation Assessment → Objective Analysis → Course of Action → Decision Output) that encodes military C2 reasoning logic as a structured, repeatable LLM prompting protocol. Second, we instantiate a PRM-inspired four-dimensional evaluation framework that scores each reasoning step on step completeness, logical coherence, decision quality, and emergence indicator—transforming qualitative C2 norms into a quantifiable process reward signal. We further conduct a Best-of-N (BoN) PRM-guided selection experiment in which five diverse CoCD responses per scenario are generated and the process-reward-maximising candidate is selected, directly demonstrating that PRM signals operate as effective inference-time guidance even without gradient-based RL training.

Throughout this paper, we use emergence to mean qualitative improvements in structured C&D reasoning that do not appear under generic prompting—specifically, the ability to synthesise novel strategies, manage contradictory information, and produce coherent multi-step decisions beyond what unstructured baselines achieve. This is weak emergence: measurable capability gains from domain-specialized scaffolding, not the appearance of entirely novel capabilities absent from the base model's pre-training distribution.

Experiments on 40 C&D scenarios of four complexity levels confirm that CoCD significantly outperforms Direct Prompting (Mann–Whitney $U = 1314$, $p < 0.0001$, Cohen's $d = 1.340$, large effect) and Standard-CoT ($p = 0.005$, $d = 0.606$, medium effect). PRM-guided BoN selection yields a further 5.8% gain over single-sample CoCD ($p < 0.001$, $d = 0.855$). A complexity-type analysis reveals that CoCD's advantage is maximised in structurally ambiguous, high-uncertainty scenarios (Level 3 gap: +0.925 points) and moderated where scenario descriptions embed explicit temporal scaffolding, providing practical guidance for deployment scope.

The core contributions of this work are: (1) the CoCD framework—the first systematic formalisation of military C2 reasoning logic as a structured LLM CoT scaffold, validated across four complexity levels; (2) the PRM-inspired process evaluation framework—demonstrating that process-level metrics expose reasoning quality differences that outcome-only evaluation masks; and (3) the BoN PRM selection experiment and complexity-type characterisation—providing direct empirical evidence for the operational utility of process reward signals. RL-based policy training on the CoCD framework is the explicit next step of this research programme.

2 Chain-of-Thought Reasoning in Large Language Models

The chain-like step-by-step thinking capability enabled by CoT prompting can trigger emergent behaviors in LLMs, allowing them to more closely mimic human reasoning processes and accomplish diverse tasks that require clear objectives and sequential steps.

2.1 CoT Prompt Engineering

The concept of CoT was first proposed by the Google Brain Team in 2022 [1], defining CoT as a series of intermediate natural language reasoning steps in the thinking process of LLMs. It was verified that CoT prompting significantly enhances the reasoning capabilities of language models, initiating the initial stage of CoT research. This technology combines the advantages of rationale-augmented training and in-context few-shot learning [2], providing the model with the ability to generate similar CoT when solving similar problems by giving prompts composed of triplets ⟨input, CoT, output⟩ as learning samples.

This enables LLMs to transition from the "fast thinking" mode for solving simple problems to the "slow thinking" mode for solving complex problems. Acquiring the ability to solve complex problems means that LLMs can further migrate and evolve into more demanding vertical domains. The chain-like thinking ability allows LLMs to more closely mimic human reasoning processes, thereby accomplishing diverse tasks with clear objectives and complete steps. However, few-shot CoT templates lack domain grounding and cannot enforce the specific inferential structure required in specialised domains: in military C2, reasoning steps must follow doctrine-aligned logic that generic "step-by-step" instructions cannot guarantee.

2.2 Built-in CoT

Built-in CoT embeds reasoning capabilities directly into LLMs through fine-tuning or training, enabling spontaneous step-by-step reasoning even for novel tasks and further amplifying emergent potential. OpenAI's o1 model exemplifies this paradigm, incorporating CoT internally to enhance the logical rationality of outputs and driving a paradigm shift for LLMs from scale expansion to reasoning expansion.

Disclosed details of the o1 model show that its built-in CoT is realized by fine-tuning external reasoning chains into the model, forming a set of reasoning tokens that represent specific CoT patterns. The model autonomously judges and invokes these tokens based on input tasks to generate structured reasoning. Early open-source replications of the o1 model, such as the g1 project [7], use dynamic CoT prompting to elicit self-reflection, enabling the model to determine whether to continue reasoning or proceed to new steps until a final answer is formed.

Qin et al. [8] argued that the o1 model's prolonged thinking process is not merely extended computation time but a human-like thorough reasoning exploration. They proposed the "Journey Learning" paradigm, which simulates human cognitive processes—including trial and error, reflection, and backtracking—and forms a Tree of Thought (ToT) [9] structure for built-in CoT. In this structure, each node represents a reasoning step, and root-to-leaf paths represent complete problem-solving trajectories, enabling the model to explore multiple reasoning strategies and solve complex exploratory or predictive tasks.

Subsequent work has further validated this paradigm: Yin et al. [10] demonstrated that combining CoT fine-tuning with MCTS-guided search and knowledge distillation enables open reasoning models to generalise across open-ended problems; Guo et al. [11] showed that pure RL training with group relative policy optimisation suffices to instil o1-level chain-of-thought reasoning in LLMs without supervised fine-tuning. These results collectively confirm that built-in CoT is a critical enabler of emergent reasoning in LLMs, and its structured, node-based reasoning pattern provides a direct design reference for the C2-specific CoCD framework proposed in this paper. Nevertheless, built-in CoT methods require substantial domain-specific training data and model retraining, making them less immediately deployable than inference-time structured prompting for resource-constrained C2 settings where retraining on classified operational data is infeasible.

2.3 Autonomous CoT

While built-in CoT endows LLMs with spontaneous reasoning capabilities, it does not guarantee the accuracy, consistency, or domain alignment of reasoning results—creating a need to explore autonomous CoT, where LLMs can self-optimize, extend, and refine reasoning chains without constant human intervention. Autonomous CoT is the key to realizing the emergence of independent C&D capabilities in LLMs, as it enables the model to adapt its reasoning to dynamic C2 scenarios without manual prompt adjustment.

Recent research has made significant progress in autonomous CoT: Ba et al. [12] proposed Automated Prompt Engineering (APE), which automatically generates and optimizes prompts to enhance model performance and generalization, laying the foundation for autonomous prompt adaptation in vertical domains; Wang et al. [13] demonstrated that LLMs can bootstrap high-quality synthetic instructions from a small set of seed tasks without human labelling (Self-Instruct), providing the foundational method for scalable, self-sustaining autonomous prompt adaptation; Hao et al. [14] proposed training LLMs to reason in a continuous latent space, maintaining multiple parallel reasoning trajectories through a dual-layer Transformer architecture that enables LLMs to autonomously extend reasoning chains and synthesise novel solutions from superposed thought trajectories—demonstrating that structured, autonomous CoT can elicit domain-specific emergence in LLMs. More broadly, recent work confirms that CoT structure is the primary driver of reasoning improvement: Li et al. [15] demonstrate that models trained on structured demonstrations generalise the reasoning scaffold itself rather than the specific content, confirming that scaffold consistency matters more than step-level detail. In structured agentic settings, Kang et al. [16] show through empirical analysis that explicit multi-step reasoning planning substantially improves task performance, validating the benefit of imposing a structured reasoning order. Furthermore, Duan et al. [17] find that uncertainty accumulates non-linearly across sequential reasoning steps in LLM agents—a key reliability challenge that domain-structured scaffolds like CoCD are specifically designed to mitigate by constraining each node's reasoning scope.

For the C&D domain, autonomous CoT requires not only the ability to self-extend reasoning chains but also the ability to align these chains with military C2 logic. This paper builds on autonomous CoT research by designing the CoCD framework, which formalizes military C&D logic into a structured, four-node autonomous reasoning chain—ensuring the domain alignment and structural stability that generic autonomous CoT methods cannot provide.

3 Process-Level Reward and Reinforcement Learning for LLM Reasoning

Reinforcement Learning (RL) is a machine learning paradigm in which an agent learns to make sequential decisions by maximising cumulative reward signals received from its environment [18]. Process-level reward modelling and RL-based optimisation are the intended training foundation of the proposed emergence mechanism, motivated by the insufficiency of outcome-only evaluation for multi-step C&D reasoning. This section reviews three lines of relevant work—RLHF, reflection mechanisms, and RL-CoT integration—and identifies the exact gap that the PRM-inspired evaluation framework proposed in this paper addresses. A Positioning paragraph at the end of this section clarifies how the proposed method relates to all prior lines.

3.1 Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) [3] is a post-training technique that incorporates human preference information into AI systems, most famously applied in ChatGPT and now a standard approach for enhancing LLM performance on open-ended and uncertain tasks. RLHF addresses

the limitation of traditional supervised training—where models only learn from labeled data—by aligning model outputs with human subjective and objective preferences through reward signals.

The RLHF framework consists of three core steps: (1) Training a base language model with strong natural language understanding and generation capabilities; (2) Collecting human preference data for model outputs and training a human preference reward model that quantifies the quality of outputs; (3) Optimizing the base model through an RL optimizer, where the model samples outputs, the reward model assigns scores, and the model updates parameters to maximize expected reward. For the C&D domain, RLHF provides a foundation for integrating military expert preferences into LLM optimization, but its focus on outcome-only reward (i.e., scoring final outputs) is insufficient for C&D tasks that require process evaluation [6]. This limitation motivates the design of the Process Reward Model (PRM) in this paper, which extends RLHF to evaluate and reward each step of the C&D reasoning process.

3.2 Reflection Mechanism

While RLHF with human-annotated data improves LLM reasoning, it suffers from high annotation costs and limited performance on highly challenging tasks—especially in vertical domains like C&D, where annotated data is scarce. Additionally, LLMs often lack the ability to self-correct flawed reasoning steps, leading to cumulative errors in long reasoning chains. To address these issues, recent research has proposed reflection mechanisms, which enable LLMs to generate feedback on their own outputs and iteratively refine reasoning—forming a self-sustaining optimization loop that reduces reliance on human annotation.

Key advances in reflection mechanisms include: Shinn et al. [19] proposed the Reflexion framework, which uses LLMs to generate detailed verbal feedback for intelligent agents based on their reasoning trajectories (rather than scalar reward values), providing comprehensive support for planning and reasoning; Madaan et al. [20] proposed SELF-REFINE, an iterative refinement method where LLMs generate initial outputs, provide self-feedback, and use this feedback to improve subsequent outputs. Reflection mechanisms are critical for the emergence of autonomous C&D capabilities in LLMs, as they enable the model to self-evaluate and correct C&D reasoning chains—an essential capability for dynamic C2 scenarios where human intervention is not always feasible. In the proposed full system, the reflection mechanism is designed to be integrated into the CoCD framework as a component of the future RL training pipeline, where the model would evaluate the rationality of each CoCD node and refine it based on RL reward signals and human feedback; this integration is part of the proposed architecture described in Section 4.1 and has not been implemented in the current work.

3.3 Integration of RL and CoT

The integration of RL and CoT is a natural and powerful paradigm for enhancing LLM reasoning, as CoT provides transparent, step-by-step reasoning structures and RL provides reward-driven optimization of these structures. This combination addresses the limitations of standalone CoT (unstable reasoning, no optimization mechanism) and standalone RL (outcome-only focus, neglect of process), enabling LLMs to generate reasoning chains that are both logically sound and optimized for domain objectives—a prerequisite for the emergence of advanced C&D capabilities.

Recent research has validated the efficacy of RL-CoT integration: Paul et al. [21] proposed the REFINER framework, which generates explicit intermediate CoT steps and uses a critic model to provide automated process feedback, enabling LLM fine-tuning without expensive human-in-the-loop data; at scale, Guo et al. [11] demonstrated that RL training on process-level reward signals enables LLMs to learn the complete thinking process behind decisions, not just the final answers—an essential property for the C&D domain, where reasoning traceability and explainability are mandatory. Recent advances in process reward modelling

extend this further: step-level verifiers that generate verification chains-of-thought (ThinkPRM [5]) and automatic conversion pipelines from outcome to process supervision [6] demonstrate that PRM-guided reasoning consistently outperforms outcome-only reward across diverse benchmarks.

For the C&D domain, RL-CoT integration must be tailored to military logical norms: reward signals must quantify the domain rationality of each reasoning step, and evaluation must focus on aligning reasoning chains with military C&D processes rather than general logical coherence alone. How this principle is instantiated in the proposed PRM-inspired framework is detailed in the Section 3.4 below.

3.4 Positioning

The CoCD framework and PRM-inspired evaluation design proposed in this paper differ from all three prior lines of work in a principled way. Unlike generic CoT prompt engineering [1,2], CoCD embeds C2-domain logic directly into the scaffold structure rather than relying on open-ended step-by-step instructions, providing structural enforcement that generic templates cannot. Unlike built-in CoT methods that train reasoning patterns into model weights [8–11], CoCD operates at inference time via prompting, making it immediately applicable to any capable LLM without retraining. Unlike existing RL-CoT integration work that evaluates only final outputs or uses general-purpose critics [21], our PRM-inspired framework evaluates each C&D reasoning step against domain-specific military norms, and demonstrates through Best-of-N selection that process signals already guide inference-time response quality—providing both the evaluation protocol and the empirical justification for future RL-PRM policy training on the CoCD framework. In terms of PRM research, Lightman et al. [4] showed that step-level verification outperforms outcome reward in mathematical reasoning; subsequent work has confirmed this finding through inference-aware fine-tuning for BoN sampling [22], compute-optimal inference strategies [23], step-level verifiers with chain-of-thought (ThinkPRM [5]), and automatic outcome-to-process supervision pipelines [6]. This paper establishes the first analogous process evaluation framework for military C2 decision-making, adapting the PRM paradigm to a structurally distinct, high-stakes domain.

4 An Emergence Mechanism in LLMs for Command and Decision-Making

Building on the theoretical foundations of CoT and RL emergence mechanisms, this paper proposes a specialized emergence mechanism for LLM-based C&D tasks. This mechanism integrates a C2-specific CoT framework (CoCD) with a PRM-inspired process evaluation framework, and incorporates contextual persistence, knowledge accumulation, and a human-in-the-loop feedback loop.

4.1 Scope of Implementation

The CoCD system architecture depicts the full target deployment, which includes RAG-based knowledge retrieval, callable API integration, MCTS-based reasoning-path search, and human-in-the-loop feedback. Of these, the currently implemented and experimentally validated components are: (1) the four-node CoCD prompting structure; (2) the PRM-inspired four-dimensional process evaluation framework; and (3) a Best-of-N PRM-guided inference-time selection study (Section 5.6). The remaining architectural components—RAG, API integration, MCTS, and RL policy training—represent design specifications for future implementation, and are described at the system level to provide context for the full research programme. This distinction follows the standard convention of systems-design papers that separate “proposed architecture” from “evaluated prototype.”

4.2 CoT Prompt Engineering in Command and Decision-Making

C&D-specific CoT prompt engineering is the foundation of the proposed emergence mechanism, as it formalizes military C&D logic into a structured, interpretable CoT format that LLMs can learn and generate autonomously. Unlike generic CoT prompt engineering, which uses open-ended step-by-step guidance, C&D CoT prompt engineering relies on domain-specialized programmed prompts that align with military C2 thinking processes. To this end, we integrate the concept of the CoCD [24] to define a standardized, four-node C&D reasoning structure, and design a precise correspondence mechanism to construct a C&D CoT training set—laying the groundwork for autonomous CoCD generation and RL optimization.

4.2.1 Form of CoT

C&D-specific CoT prompting requires designing military domain-aligned prompt templates that guide LLMs to generate reasoning chains consistent with commander decision-making logic, avoiding non-alignment with military backgrounds and low reasoning quality caused by generic prompts. The primary form of C&D CoT prompting is instructional prompting, where military C&D tasks are decomposed into sequential thinking step nodes, and clear command prompts guide LLMs to generate step-by-step reasoning and decisions.

To standardize this process, we propose the CoCD framework—a four-node C2-specific CoT structure that aligns with the commander's thinking process of information collection, objective analysis, action planning, and decision execution [24]. The CoCD framework defines each reasoning step in the C&D process as a node, with the CoT composed of four sequentially linked nodes:

- **Situation Assessment:** Parse task input, retrieve relevant C2 knowledge, and analyze the current operational situation;
- **Objective Analysis:** Define core operational objectives, identify constraints and priorities, and trade off competing goals;
- **Course of Action:** Generate feasible operational action plans, evaluate the pros and cons of each plan, and screen for optimal options;
- **Decision Output:** Form a final executable decision, provide explicit justification, and generate digital C&D products (e.g., operational orders).

In the CoCD framework, each node has three core components (Fig. 1): (1) Callable API codes for algorithm tools and knowledge corpus access; (2) Domain knowledge support from a C2 corpus and Retrieval-Augmented Generation (RAG) technology; (3) Contextual inheritance where the content generated by the previous node serves as the initial reference and basis for the next node. This structure ensures the contextual persistence, logical coherence, and domain alignment of the C&D reasoning process—critical properties for the emergence of reliable C&D capabilities in LLMs. The CoCD framework transforms the unstructured C&D reasoning of commanders into a structured, repeatable LLM reasoning process, enabling quantitative evaluation and RL optimization of each step.

4.2.2 The Construction of CoT Training Set

To train LLMs to generate autonomous CoCD reasoning chains, we design a precise correspondence mechanism to construct a high-quality C&D CoT training set—one where each training sample is a structured CoCD chain that precisely corresponds to a C&D task input. This mechanism ensures that the training set is aligned with military C2 logic and provides sufficient supervision for LLMs to learn CoCD generation.

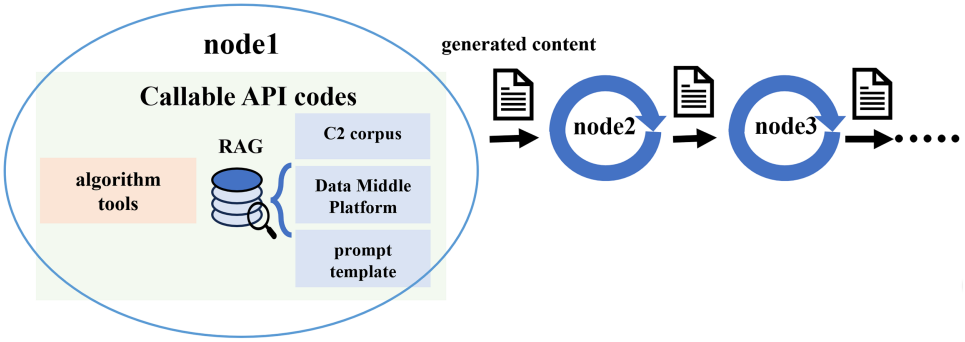

Figure 1: The form of CoT in the field of command and decision-making.

The precise correspondence mechanism consists of four sequential steps (Fig. 1):

- **Task Semantic Parsing:** The LLM parses the C&D task requirements input by the commander, identifying key entities, operational objectives, constraints, and environmental factors;
- **Domain Knowledge Retrieval:** Based on the parsed key elements, RAG technology retrieves relevant background knowledge, operational constraints, and empirical references from a specialized C2 corpus;
- **Prompt Template Extraction:** A predefined template library extracts a CoCD-aligned prompt template based on task characteristics (e.g., single-objective, multi-objective, uncertain), which integrates the CoCD four-node structure, task details, and retrieved corpus information;
- **Structured CoCD Generation:** Guided by the prompt template, the LLM generates a structured CoCD reasoning chain that precisely corresponds to the task input, with each node containing domain-rational content and logical connections to adjacent nodes.

All high-quality CoCD chains generated via this mechanism are accumulated as a C&D CoT training set, which is fed into the LLM through prompt engineering and parameter fine-tuning. This training process endows the base LLM with basic C&D reasoning capabilities and establishes the initial CoCD generation pattern—an essential prerequisite for subsequent RL optimization and the emergence of autonomous C&D capabilities.

4.3 Application of Reinforcement Learning in the C&D Process

Reinforcement Learning is the intended core optimization driver of the proposed emergence mechanism, designed to be applied throughout the CoCD generation process to refine reasoning quality and enable autonomous self-improvement. This subsection describes both the implemented evaluation components and the proposed RL training pipeline.

Contextual Continuity. In the chain-of-thought format within the command and decision-making domain, the output of each node serves as the input for subsequent nodes. This chain-like structure constructs a continuously evolving, highly information-correlated contextual background, ensuring the coherence and logical consistency of the reasoning process, and also providing data references for the reward algorithm of the next node. Contextual continuity transforms the CoCD framework into a Markov Decision Process (MDP), where each node represents a state, the transition from one node to the next represents an action, and the reward signal quantifies the quality of the state and action. This MDP formulation lays the mathematical foundation for the PRM.

PRM-Inspired Process Evaluation. Unlike traditional reward functions that solely evaluate the correctness of final decisions, the PRM-inspired evaluation framework assesses each step of reasoning in the chain

of thought across four dimensions: step completeness (coverage of required reasoning steps, 0–10), logical coherence (causal soundness and flow, 0–10), decision quality (operational rationality of the final decision, 0–10), and emergence indicator (reasoning synthesis beyond template matching, 0–10). The composite score is the unweighted mean of the four dimensions. This four-dimensional framework operationalizes process-level reward signals, enabling comparison between PRM-guided and ORM-guided (outcome-only) selection as demonstrated in Section 5.6.

RL-Based PRM Training. Building on the MDP formulation, the intended next phase of this research is to use the PRM evaluation framework as a reward function for RL policy training. The MDP formulation of CoCD nodes supports a GRPO (Group Relative Policy Optimization) or PPO optimization objective, where the policy is the LLM generating CoCD steps and the reward is the per-node PRM score. Combined with Monte Carlo Tree Search for path exploration, this pipeline is designed to optimize the full C&D reasoning trajectory rather than the final decision alone. Technical specifications (reward function formulation, optimization algorithm, training protocol) are described at the design level; no RL training has been performed in this submission.

Keeping Humans on the Loop. Commanders always retain the highest authority to evaluate the output content of LLMs, ensuring ethical oversight of AI-assisted decisions. The results of human-in-the-loop evaluations are collected to form a feedback loop, which is used to dynamically optimize prompt engineering and knowledge retrieval strategies, enabling the system to continuously learn from feedback provided by human experts. All verified high-quality chains of thought and their results will be fed back into the command and control corpus or prompt template library, endowing the system with the capability for self-enrichment and evolution.

4.4 Emergence Mechanism Architecture of LLMs for C&D

As noted in the Scope of Implementation (Section 4.1), the architectural components described below represent design specifications for the full target system; only the four-node CoCD prompting structure, the PRM-inspired evaluation framework, and the BoN selection study have been implemented and experimentally validated. By establishing a quantitative evaluation mechanism for the reasoning process and applying reinforcement learning reward signals to each reasoning step in the chain of thought, autonomous chain-of-thought generation in the field of command and decision-making is designed to be achieved, thereby enabling reliable and controllable intelligent emergence in command and decision-making tasks.

Combining the training methods of LLMs for command and decision-making and workflow LLMs [24], a basic framework for LLMs with built-in command and decision-making chains is constructed, as shown in Fig. 2.

Model Training. Utilizing the training set of chains of thought generated through a precise correspondence mechanism, the chains are integrated into the LLMs via prompt engineering, parameter fine-tuning, and other methods, enabling the foundational LLMs to possess basic reasoning capabilities in the field of command and decision-making.

Generating Chain-of-Thought. Applying the tree-structured exploration format of the Tree of Thoughts (ToT) approach, together with a process reward model and search algorithms such as Monte Carlo Tree Search (MCTS), multiple reasoning paths can be generated and compared. CoT Self-Consistency (CoT-SC) is a decoding and aggregation strategy that samples diverse reasoning paths and selects the most consistent answer, typically by majority voting; it is not itself an evaluation metric. These complementary mechanisms can be used to generate a high-quality chain of thought.

Humans on the Loop. Commanders assess the quality of the generated chains of thought and decide whether to adopt them or reconsider the output. The manually evaluated chains of thought are written back

to a specialized knowledge corpus for data augmentation and further reinforcement learning. Through this feedback iteration process, the generated chains of thought gradually align with the task scenarios and meet the commanders' requirements, achieving autonomous chain-of-thought prompting capabilities in the field of command and decision-making.

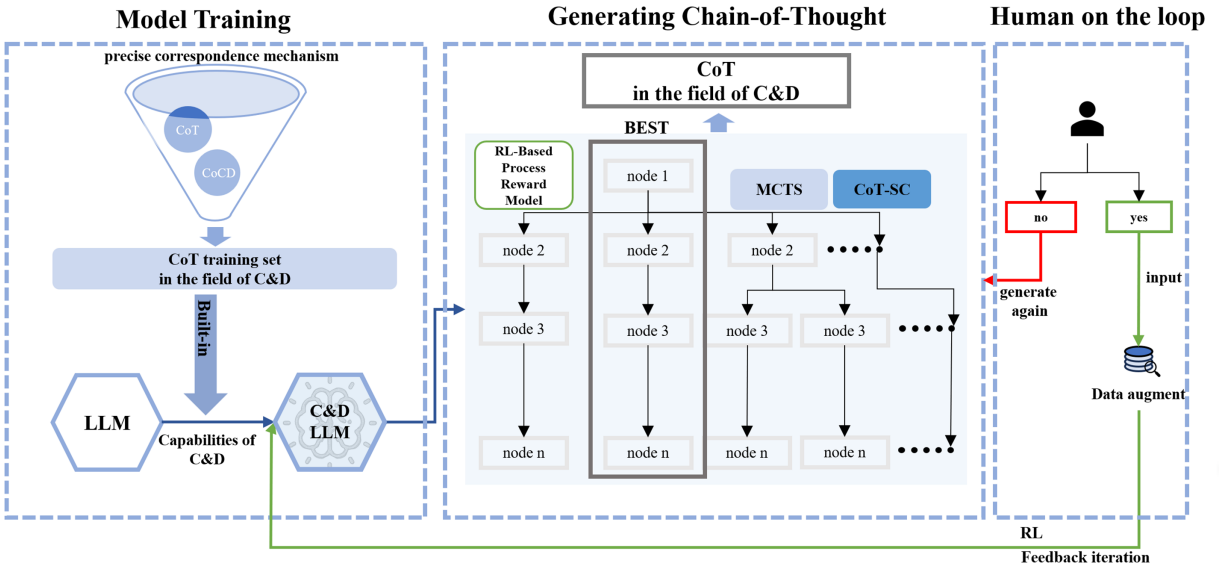

Figure 2: An emergence mechanism architecture of LLMs for command and decision-making.

5 Experiment

To empirically validate the effectiveness of the proposed emergence mechanism and the Chain of Command and Decision (CoCD) framework for large language model (LLM)-driven command and decision-making (C&D) tasks, a controlled comparative experiment was designed and conducted across 40 C&D scenarios with graded complexity. This experiment benchmarks the performance of the CoCD framework against two widely adopted prompting baselines: Direct Prompting, which provides no explicit reasoning guidance for the model, and Standard Chain-of-Thought (Standard-CoT) prompting, which leverages the generic "Let's think step by step" prompt to elicit open-ended sequential reasoning. In addition, a Best-of-N PRM-guided selection experiment (Section 5.6) and a cross-model judge validation experiment (Section 5.7) are conducted to provide direct evidence for PRM utility and evaluation robustness respectively.

5.1 Experimental Setup

DeepSeek-V3 was selected as the test model for its strong complex reasoning capabilities and the absence of pre-training on military C&D domain data, eliminating domain fine-tuning bias and ensuring experimental neutrality. The LLM-as-Judge paradigm was employed for performance evaluation, with DeepSeek-V3 scoring all three prompting strategies under uniform criteria. To assess potential self-consistency bias, an independent cross-model evaluation was conducted using Qwen3 as judge (Section 5.7).

Three prompting strategies were designed for the experiment, as summarised in Table 1: Direct Prompting, as the naive baseline, required the model to generate C&D decisions directly without intermediate reasoning steps; Standard-CoT, the generic structured reasoning benchmark, used canonical step-by-step guidance to elicit open-ended reasoning chains; the CoCD strategy leveraged the domain-specialized four-node instructional prompt template, guiding the model to generate structured reasoning in the sequence of

Situation Assessment → Objective Analysis → Course of Action → Decision Output, aligned with military C&D logical norms. The scenario design and evaluation metrics are reported in Tables 2 and 3, respectively.

Table 1: Prompting strategies.

Strategy	Description
Direct	Baseline. No reasoning guidance. Model answers directly.
Standard-CoT	"Let's think step by step"—standard chain-of-thought prompting.
CoCD	Domain-specific four-node framework: Situation Assessment → Objective Analysis → Course of Action → Decision Output.

Table 2: Scenario design.

Level	Category (n = 10)	Description
Level 1	Single-objective	Simple decisions with one clear objective
Level 2	Multi-objective	Trade-offs between competing goals
Level 3	Uncertainty	Decisions under incomplete/conflicting information
Level 4	Sequential	Multi-phase decisions with cascading dependencies

Table 3: Evaluation metrics.

Metric	Description
Step Completeness	Coverage of all required decision-making steps
Logical Coherence	Causal soundness and logical flow between steps
Decision Quality	Operational soundness and justification of the final decision
Emergence Indicator	Reasoning beyond simple pattern matching; novel synthesis
Composite	Mean of the four metrics above

Forty C&D scenarios were stratified into four complexity levels (10 scenarios per level), mirroring the characteristics of real-world military C&D tasks: Level 1 (Single-objective) for low-complexity tasks with a single clear operational objective; Level 2 (Multi-objective) for moderate-complexity tasks requiring trade-offs between competing goals; Level 3 (Uncertainty) for high-complexity tasks under incomplete/conflicting operational information; Level 4 (Sequential) for the highest-complexity multi-phase tasks with cascading step dependencies. Full scenario texts are provided in Appendix A.

A comprehensive evaluation system was defined, including four primary 0–10 scale metrics and one composite metric (unweighted mean of the four), capturing both process and outcome quality of C&D reasoning: step completeness (coverage of required reasoning steps), logical coherence (causal soundness and flow between steps), decision quality (operational and military rationality of final decisions), and emergence indicator (novel reasoning synthesis beyond pattern matching). The composite metric served as the core indicator for overall performance comparison of the three strategies.

5.2 Overall Performance Comparison

The CoCD framework achieved the highest composite performance score across all 40 scenarios, reaching 9.00 ± 0.45 (DeepSeek-V3 judge)—3.8% higher than Standard-CoT (8.67 ± 0.62) and 7.8% higher

than Direct Prompting (8.35 ± 0.51). Statistical testing confirms that these differences are robust: Mann–Whitney $U = 1314$, $p < 0.0001$, Cohen's $d = 1.340$ (large effect) for CoCD vs. Direct; $U = 1062$, $p = 0.005$, $d = 0.606$ (medium effect) for CoCD vs. Standard-CoT; and $U = 1076$, $p = 0.004$, $d = 0.556$ (medium effect) for Standard-CoT vs. Direct. The overall comparison is shown in Fig. 3.

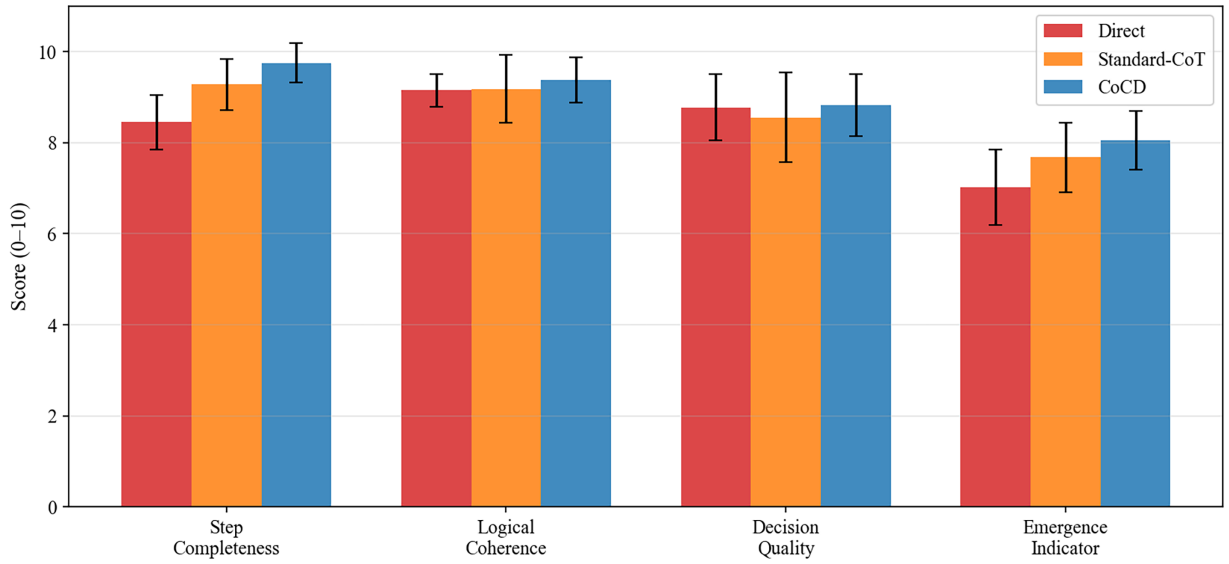

Figure 3: Overall performance comparison across prompting strategies.

On the step completeness dimension, CoCD achieved the highest score of 9.75 ± 0.44, demonstrating the four-node structure's robust ability to enforce full coverage of required C&D reasoning steps. All three strategies exhibited high logical coherence (9.15–9.38), indicating that structured prompting does not compromise causal soundness. For the emergence indicator, CoCD achieved 8.05 ± 0.64, 14.5% higher than Direct Prompting (7.03 ± 0.83), confirming that the CoCD framework effectively elicits novel reasoning synthesis.

5.3 Performance by Task Complexity

Analysis of the 40-scenario dataset reveals a nuanced complexity-type effect rather than a simple monotonic scaling. CoCD's composite score advantage over Direct Prompting widened progressively from Level 1 to Level 3: +0.675 points at Level 1 (CoCD: 8.875 vs. Direct: 8.200), +0.875 at Level 2 (9.225 vs. 8.350), and a maximum +0.925 at Level 3 (9.125 vs. 8.200). However, the gap narrowed sharply to +0.125 at Level 4 (8.775 vs. 8.650). The complexity trend is visualised in Fig. 4.

Importantly, this narrowing is not caused by CoCD degrading at Level 4—CoCD's absolute score is essentially stable (Level 3: 9.125; Level 4: 8.775, $\Delta = -0.35$). Rather, the gap narrows because Direct Prompting improves substantially at Level 4 (Level 3: 8.200→ Level 4: 8.650, $\Delta = +0.450$). One plausible interpretation is that Level 4 scenarios are multi-phase sequential planning tasks containing explicit temporal and structural scaffolding in the problem description itself (e.g., "Phase 1 (0–24h):", cascade trigger conditions). This embedded structure may provide a natural organising scaffold that even unguided Direct Prompting can exploit, reducing CoCD's incremental structural advantage. Level 3 uncertainty scenarios, by contrast, present probabilistic ambiguity with no inherent structural cues—precisely the condition where CoCD's imposed four-node framework delivers the largest benefit. A systematic content analysis of all 40 scenario descriptions to formally quantify embedded scaffolding is left to future work.

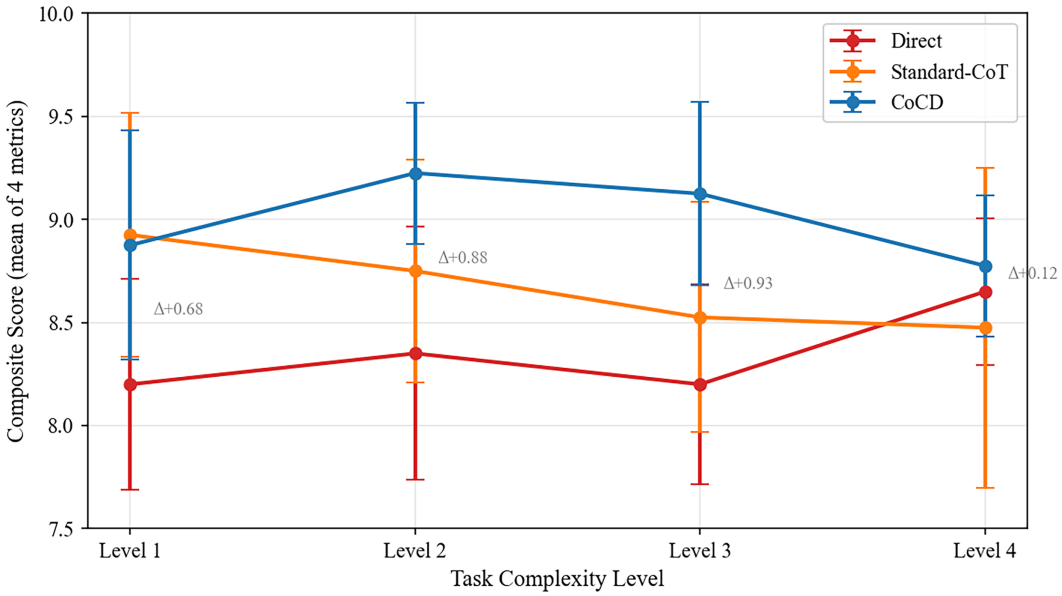

Figure 4: Emergence of CoCD advantage with increasing task complexity.

CoCD's structural advantage is therefore most pronounced in structurally ambiguous, high-uncertainty environments, and is moderated in scenarios where the task description itself provides a temporal or phase scaffold. This complexity-type characterisation is more accurate than the original monotonic non-linearity claim, and provides actionable guidance for the deployment scope of structured CoT frameworks in military C2 systems: CoCD delivers the greatest incremental value for unstructured, uncertainty-heavy scenarios typical of real-world operational intelligence tasks.

In contrast, Standard-CoT exhibited a declining performance trend with increasing complexity: highest at Level 1 (8.925) but dropping to 8.750 at Level 2, 8.525 at Level 3, and 8.475 at Level 4, even underperforming Direct at Level 4. This divergence confirms that the CoCD framework's four-node design sustains reasoning quality under complexity conditions that overwhelm open-ended Standard-CoT.

The emergence indicator analysis further verified that CoCD is more effective at triggering domain-specific emergent reasoning than the two baselines, with an overall score of 8.05 ± 0.63—the highest among the three strategies. CoCD's emergence indicator peaked at Level 2 (8.50), outperforming Standard-CoT (7.60) and Direct Prompting (7.00), as multi-objective tasks demand novel compromise synthesis between competing goals, a typical manifestation of emergent reasoning that CoCD's structured framework is designed to elicit. The complete complexity-level scores are reported in Table 4.

For Level 3 (uncertainty) tasks, CoCD maintained a high emergence indicator score of 8.20, exceeding Standard-CoT (7.80) and Direct Prompting (6.90) by 0.40 and 1.30 points respectively, demonstrating its ability to drive adaptive novel reasoning under incomplete information. For Level 4 (sequential) tasks, CoCD's emergence indicator (7.80) matched Standard-CoT (7.80), with Direct Prompting close behind (7.70)—consistent with the complexity-type effect: when scenario descriptions embed explicit temporal scaffolding, all three strategies converge in their ability to elicit structured synthesis. For Level 1 (single-objective) tasks, CoCD scored 7.70 vs. Standard-CoT (7.50) and Direct Prompting (6.50), reflecting modest structured overhead relative to simple tasks but a clear advantage over the unguided baseline.

Table 4: Composite score by complexity level (mean, DeepSeek-V3 judge).

Level	CoCD	Standard-CoT	Direct	Δ (CoCD–Direct)
Level 1 (Single-obj.)	8.875	8.925	8.200	+0.675
Level 2 (Multi-obj.)	9.225	8.750	8.350	+0.875
Level 3 (Uncertainty)	9.125	8.525	8.200	+0.925
Level 4 (Sequential)	8.775	8.475	8.650	+0.125

5.4 PRM vs. ORM Evaluation

To validate the PRM design's rationality, the experiment compared ORM (proxied by decision quality alone, selected because it most directly corresponds to the traditional outcome reward definition of evaluating only the final answer) and PRM (proxied by the composite metric) for Standard-CoT and CoCD, quantifying ORM overestimation—cases where a high final decision score masks flawed intermediate reasoning, a critical issue for the C&D domain's demand for reasoning traceability. The PRM–ORM comparison is shown in Fig. 5.

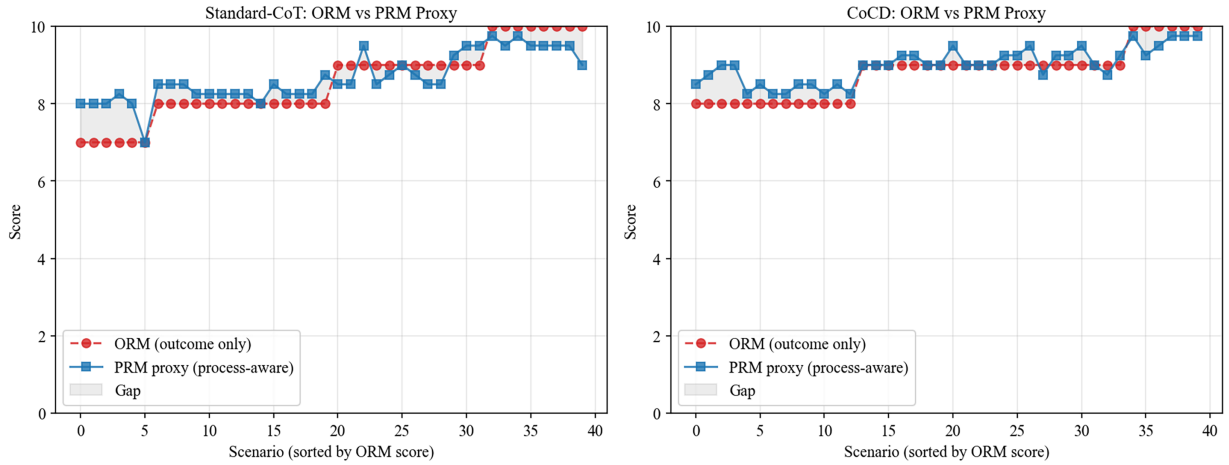

Figure 5: Process reward model vs. outcome reward model.

Results showed that CoCD significantly reduced the frequency and magnitude of ORM overestimation compared to Standard-CoT: ORM overestimated performance in 10 of 40 scenarios (25%) for Standard-CoT, with a mean overestimation gap of 0.60 points, while for CoCD, overestimation occurred in only 6 of 40 scenarios (15%) with a smaller 0.50-point mean gap. These findings confirm that ORM is an insufficient evaluation mechanism for C&D tasks, as it ignores process-level flaws, and validate the necessity of PRM's process-level optimization. The CoCD framework's structured reasoning process minimizes disconnect between reasoning quality and final decisions, making PRM a more robust evaluation and optimization foundation for LLM-driven C&D tasks.

5.5 Performance by Scenario Category

Fig. 6 presents composite scores broken down by scenario category. The pattern confirms the complexity-type characterisation established in Section 5.3: CoCD achieves the highest composite across all four categories, with the largest margins in uncertainty and multi-objective scenarios (+0.925 and +0.875 over Direct respectively), and the narrowest margin in the sequential category where embedded scaffolding benefits all strategies.

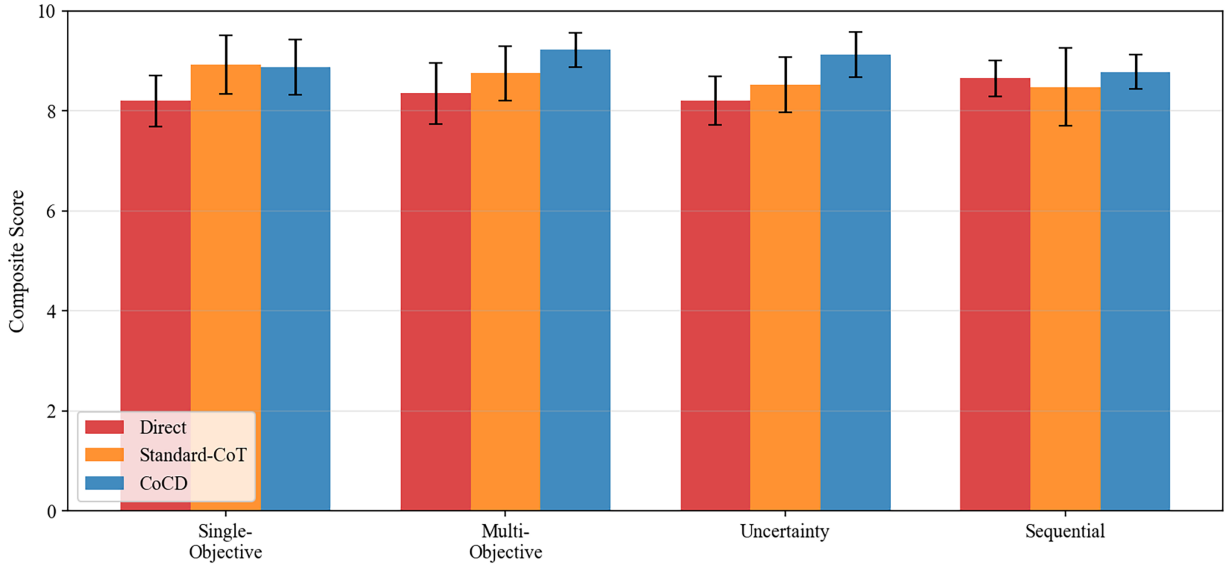

Figure 6: Performance by scenario category.

5.6 PRM-Guided Best-of-N Selection

To provide direct empirical evidence that the PRM evaluation framework operationally guides response quality, a Best-of-N (BoN) selection experiment was conducted. For each of the 40 C&D scenarios, five diverse CoCD responses were generated at temperature 0.7. $N = 5$ was selected as a computationally tractable starting point that balances candidate diversity against evaluation cost; temperature sampling at 0.7 provides sufficient output diversity within a single model without requiring multiple decoding strategies. All five candidates were scored by the PRM (four-metric composite) using Qwen3 [25] as an independent judge, and two selection criteria were compared. The selection-performance comparison is shown in Fig. 7.

- **CoCD-BoN-ORM:** select the candidate with the highest decision_quality score alone (outcome reward signal).
- **CoCD-BoN-PRM:** select the candidate with the highest four-metric composite score (process reward signal).

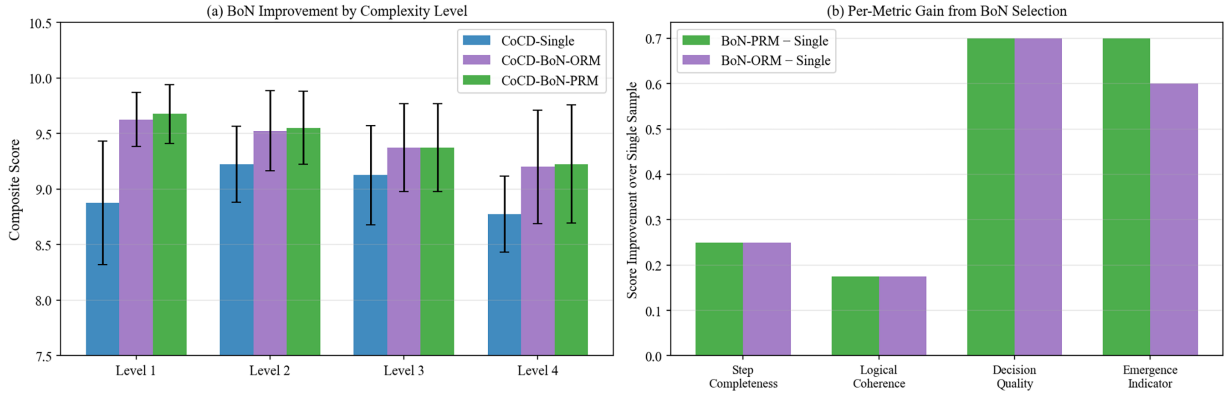

Figure 7: Best-of-N PRM vs. ORM selection performance (Qwen3 judge, $N = 5$, 40 scenarios).

Results are summarised in Table 5. CoCD-BoN-PRM achieved a mean composite of 9.456 ± 0.410, representing a 5.8% improvement over single-sample CoCD (8.938 ± 0.741; Mann–Whitney $U = 1206$, $p < 0.001$, Cohen's $d = 0.855$, large effect). CoCD-BoN-PRM marginally outperformed CoCD-BoN-ORM ($\Delta = 0.025$ composite points). A Wilcoxon signed-rank test on the 40 paired scenario scores yields $W^{+} = 10$, $p = 0.068$ (two-tailed), indicating that the difference does not reach conventional statistical significance at this sample size. PRM scores exceeded ORM scores in 4 of 40 scenarios and were equal in the remaining 36, reflecting that both criteria selected the same candidate in the majority of cases. The directional consistency of the advantage warrants investigation at larger N.

Table 5: Best-of-N Selection Results ($N = 5$, 40 scenarios). All composite scores are from the Qwen3 judge; the DeepSeek-V3-judged CoCD-single composite is 9.00 ± 0.45 (Section 5.2).

Condition	Mean Composite	vs. CoCD-Single
CoCD-single (Qwen3 judge)	8.938 ± 0.741	Baseline
CoCD-BoN-ORM	9.431 ± 0.403	+5.5%
CoCD-BoN-PRM	9.456 ± 0.410	+5.8%, $p < 0.001$, $d = 0.855$

This experiment directly implements the inference-time PRM selection paradigm [4,22,23,26], demonstrating that PRM guidance is operationally effective for C&D response selection even without gradient-based RL training. The result establishes an empirical foundation for future RL-PRM policy training: if process-aware signals already improve selection quality at inference time, RL training on these signals is expected to further improve the generation quality of the base policy.

5.7 Cross-Model Judge Validation

To assess potential self-consistency bias from using DeepSeek-V3 as both generator and judge, all 40-scenario responses were independently re-evaluated using Qwen3 [25] (Alibaba Qwen series, accessed via SiliconFlow API)—an independently developed model with no architectural overlap with DeepSeek-V3.

Cross-judge correlation across 120 paired scenario–strategy scores: Pearson $r = 0.389$, Spearman $\rho = 0.453$. Both judges agree on the ranking of strategies (CoCD > Standard-CoT > Direct), confirming that the qualitative finding is model-agnostic. The Qwen3-judged CoCD composite (8.938 ± 0.741) is comparable to the DeepSeek-V3-judged value (9.000 ± 0.447), with no evidence of self-favouring by the DeepSeek-V3 judge. The moderate score-level correlation reflects that Qwen3 assigns more variable absolute scores (Direct std: 1.368 vs. DeepSeek-V3's 0.509), suggesting stricter score differentiation rather than systematic inflation by the DeepSeek-V3 judge.

These results support the robustness of the main experimental findings while honestly acknowledging that absolute LLM-as-Judge scores are judge-dependent. Future work should incorporate human expert evaluation to calibrate both judge models against ground-truth C&D quality assessments.

6 Conclusion and Prospects

This paper proposes and empirically validates an LLM emergence mechanism for command and decision-making (C&D) tasks built on two implemented components: a domain-specialized Chain of Command and Decision (CoCD) framework and a PRM-inspired four-dimensional process evaluation framework. Within the scope of a controlled prompt-based evaluation on 40 simulated C&D scenarios, CoCD-structured reasoning consistently produced higher process quality and decision scores than Direct

Prompting and Standard-CoT baselines (CoCD vs. Direct: $p < 0.0001$, Cohen's $d = 1.340$, large effect). PRM-guided Best-of-N selection further improved performance by 5.8% over single-sample CoCD ($p < 0.001$, $d = 0.855$), providing direct empirical evidence that process-aware reward signals operationally guide response quality at inference time. Analysis across complexity levels reveals a complexity-type effect: CoCD delivers the greatest advantage in high-uncertainty, structurally ambiguous scenarios (Level 3: +0.925 gap), while explicit temporal structure in sequential scenarios allows Direct Prompting to partially close the gap.

These results provide empirical support for domain-specialized structured reasoning and process-level evaluation as foundations for future RL-based C&D capability development in LLMs, but do not constitute evidence of deployment-ready autonomous C&D systems. The gap between controlled evaluation and operational deployment remains substantial.

Limitations. The evaluation relies on the LLM-as-Judge paradigm; cross-model validation (Pearson $r = 0.389$, Spearman $\rho = 0.453$) confirms consistent strategy rankings but moderate score-level agreement. The relatively low correlation—even at the rank level—reflects an inherent limitation of the LLM-as-Judge paradigm: absolute scores are judge-dependent, and the extent to which any single judge's scores constitute ground-truth quality remains uncertain. Additionally, the PRM four-dimensional metrics are each scored by a single judge without inter-rater reliability evaluation at the individual dimension level. Future work should incorporate human expert evaluation and standard inter-rater reliability measures (e.g., ICC or Krippendorff's α) to calibrate judge performance against ground-truth C&D quality assessments. No RL policy training has been performed; the PRM framework described in Section 4.3 and the BoN experiment provide inference-time evidence only. A formal component-level ablation of the full proposed architecture (RAG, MCTS, human-in-the-loop) is deferred to future work pending implementation of those components. Scenarios are designed to represent real-world C&D tasks but are simulated; real operational validation remains necessary. The cross-judge score relationship is visualised in Fig. 8.

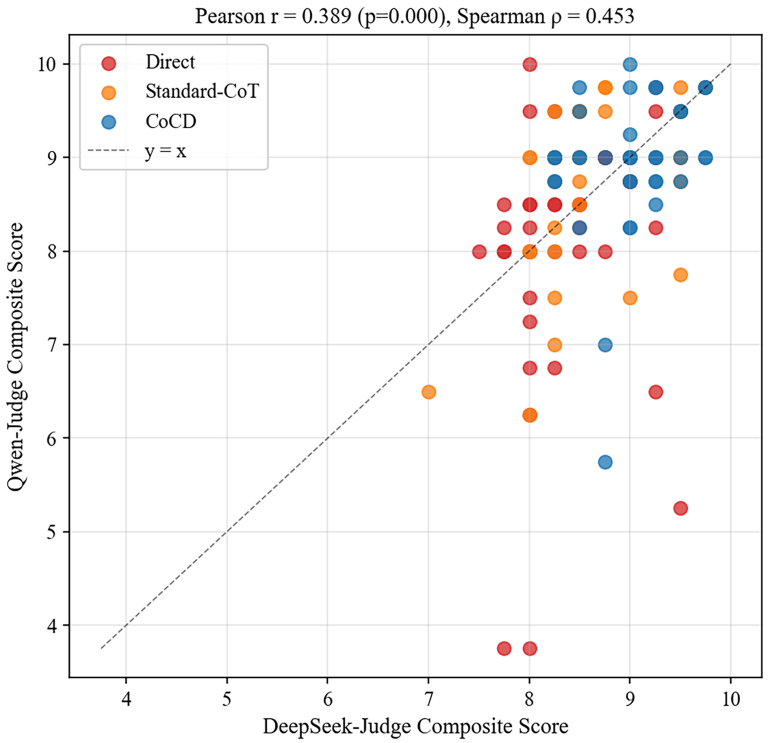

Figure 8: Cross-judge score correlation: DeepSeek-V3 judge vs. Qwen3 judge (120 paired scenario–strategy scores).

In the future, group decision-making under multi-agent collaboration could be incorporated to address even more complex task scenarios [27], and the PRM-inspired evaluation framework described here provides a natural reward signal for RL policy training on the CoCD framework.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Yazhi Zheng; Methodology, Yazhi Zheng; Writing—original draft preparation, Yazhi Zheng; Supervision, Xiaolong Cui; Project administration, Xiaolong Cui; Writing—review and editing, Xiaolong Cui and Xuanzhu Sheng; Investigation, Xin Wang; Validation, Xin Wang; Formal analysis, Xin Wang; Data curation, Xuanzhu Sheng; Visualization, Xuanzhu Sheng. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, X.C., upon reasonable request, subject to institutional and security review.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A All 40 C&D Evaluation Scenarios

Table A1 lists all 40 command and decision-making scenarios used in the experiments. Each scenario is classified by complexity level and category.

Table A1: All 40 C&D scenarios (Appendix A).

ID	Lvl	Category	Scenario Summary
1	1	Single-obj.	Enemy supply convoy intercept: 3 vehicles at 15 km/h, strike team 20 min away, intercept in 40 min. Decide: intercept now?
2	1	Single-obj.	Base resupply: supplies for 3 days, convoy 6 h away, storm in 4 h lasting 8 h. Depart now or wait?
3	1	Single-obj.	Relay station power failure: 2 h on backup. Team A 90 min (certain fix); Team B 45 min (60% fix). Which team?
4	1	Single-obj.	Perimeter defense: 3 platoons, north approach 70% attack probability, east 30%. Allocate forces.
5	1	Single-obj.	Medevac: 4-person capacity, 2 critical, 3 serious, 1 minor. Single trip possible. Prioritize.
6	2	Multi-obj.	Bridge assault by dawn: Option A 80%/30% casualties; Option B 60%/5%; Option C 95%/0% (4 h delay).
7	2	Multi-obj.	Cyber intrusion on 3 systems simultaneously; resources to fully defend 2; partial defense reduces each by 40%.
8	2	Multi-obj.	Advance 50 km in 12 h: Route A direct/minefield; Route B safe/12.5 h; Route C 65 km/2-h river crossing.

(Continued)

Table A1 (continued)

ID	Lvl	Category	Scenario Summary
9	2	Multi-obj.	Enemy command post near civilian hospital: artillery 95%/high collateral; SF 70%/low; cyber 50%/zero.
10	2	Multi-obj.	Allocate 100 budget units across intelligence, firepower, logistics with minimum viable thresholds.
11	3	Uncertainty	Contradictory recon: Source A (70%) reports 500 troops; Source B (85%) reports 150; Source C confirms A. Attack in 2 h?
12	3	Uncertainty	Drone spots possible missile launcher: 60% real, 30% decoy, 10% civilian. One precision strike available.
13	3	Uncertainty	Forward unit comms cut 4 h: 50% achieved objective, 30% engaged, 20% lost. Send reserve or hold?
14	3	Uncertainty	Amphibious landing: tomorrow 40% calm/25% rough; delay 24 h gives 70% calm but 60% enemy reinforce risk.
15	3	Uncertainty	Captured document: 65% genuine, 25% disinformation, 10% outdated. Delay 6 h to analyze or proceed?
16	4	Sequential	3-phase 72-h operation: allocate 60%/20%/20%; Phase 1 failure reduces Phase 2 success by 40%. Plan transitions.
17	4	Sequential	Week-long counter-insurgency: intelligence gathering→cordon→raids→handover; civilian cooperation degrades with aggression.
18	4	Sequential	6-h crisis with three escalation paths (40% diplomatic/35% limited military/25% full conflict). Decision tree at 0, 2, 4 h.
19	4	Sequential	Joint 96-h operation: Army/Navy/Air Force with different response times; 5 sequential objectives. Coordination framework.
20	4	Sequential	Evolving situation T = 0 h to T = 3 h with four sequential intel updates and force splits. Adaptive decision process.
21	1	Single-obj.	Artillery fire support requested; friendly platoon enters danger zone in 12 min, radio intermittent. Fire now or wait?
22	1	Single-obj.	Night patrol route: Route X (4 km, village, insurgent risk) vs. Route Y (7 km, 5.5 h, open terrain). 6-h window.
23	1	Single-obj.	Withdrawal: 8 injured civilians (3 critical), vehicle capacity for 4 extras, medical facility 15 min opposite direction.
24	1	Single-obj.	Two platoons need resupply: Platoon A 20% ammo/defensive, Platoon B 15%/active firefight. One vehicle.

(Continued)

Table A1 (continued)

ID	Lvl	Category	Scenario Summary
25	1	Single-obj.	Observation post: Position Alpha 270° view/2 personnel vs. Position Beta 180°/1 personnel. 18-h mission.
26	2	Multi-obj.	Enemy logistics hub: airstrike (80% supply denial) vs. capture intact (25% casualty risk) vs. road block (temporary).
27	2	Multi-obj.	Bridge fate: destroy (cuts own resupply 10 days) vs. defend (commits reserve) vs. delay demolition (enemy capture risk).
28	2	Multi-obj.	Intelligence asset extraction: helicopter (10 min, alerts garrison 20 min away) vs. ground (45 min, silent, 8 personnel).
29	2	Multi-obj.	EW jamming choice: full-spectrum (high effect, disrupts own comms, reveals capability) vs. selective (60%) vs. none (preserves surprise).
30	2	Multi-obj.	Medical resource allocation: 2 surgeons, supplies for 8 surgeries, 14 soldiers + 6 civilians needing surgery, 2nd wave in 8 h.
31	3U	Uncertainty	Grid 447: aerial imagery (trucks, 80%), SIGINT (field hospital, 65%), HUMINT (ammo depot, 50%). Strike or not?
32	3	Uncertainty	Night assault: 30% thermal degradation, 55% fog probability, 70%/40% confidence on enemy positions. Proceed or delay?
33	3	Uncertainty	Counter-battery: 75% location confidence, 3 options (precision arty/standard arty/air support 25 min). Civilian risk high.
34	3	Uncertainty	Ceasefire declared 6 h ago: 60% genuine/40% deception. Maintain posture, advance, or request verification?
35	3	Uncertainty	Company silent 90 min: 45% equipment damage, 25% needs reinforcement, 20% deliberate silence, 10% overrun. 30-min window.
36	4	Sequential	4-day deception operation before main assault with Day 1 feint (70% success), Day 2–3 reconnaissance gate, Day 4 assault.
37	4	Sequential	4-day evacuation of 200 personnel: helicopter (40/sortie, 2/day, 70% weather) + ground convoy (60/convoy, escort required).
38	4	Sequential	96-h FOB establishment: 4 phases with air defense installation window (35% attack risk) as critical gate before Phase 3.
39	4	Sequential	Coalition operation: Nation A (heavy armor, aggressive ROE), Nation B (SOF, 2-h authorization delay), Nation C (aviation, 4-h planning).
40	4	Sequential	6-h defensive battle with sequential intel updates at T = 0, 1.5, 3, 4.5 h. Reserve commitment decision framework.

References

1. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing System; 2022 Nov 28–Dec 9; New Orleans, LA, USA. Red Hook, NY, USA: Curran Associates Inc.; 2022. p. 24824–37.
2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 1877–901.
3. Christiano P, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 4302–10.
4. Lightman H, Kosaraju V, Burda Y, Edwards H, Baker B, Lee T, et al. Let's verify step by step. In: Proceedings of the International Conference on Learning Representations; 2024 May 7–11; Vienna, Austria.
5. Khalifa M, Agarwal R, Logeswaran L, Kim J, Peng H, Lee M, et al. Process reward models that think. arXiv:2504.16828. 2025.
6. Zheng C, Zhu J, Ou Z, Chen Y, Zhang K, Shan R, et al. A survey of process reward models: from outcome signals to process supervisions for large language models. arXiv:2510.08049. 2025. doi:10.48550/arxiv.2510.08049.
7. Qin Y, Li X, Zou H, Liu Y, Gu S, Ding L, et al. O1 replication journey [Internet]. GitHub; 2024 [cited 2026 Aug 10]. Available from: https://github.com/GAIR-NLP/O1-Journey.
8. Qin Y, Li X, Zou H, Liu Y, Gu S, Ding L, et al. O1 replication journey: a strategic progress report—Part 1. arXiv:2410.18982. 2024.
9. Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook, NY, USA: Curran Associates Inc.; 2023. p. 11809–22.
10. Yin H, Zhao Y, Wu M, Ni X, Zeng B, Wang H, et al. Marco-o1 v2: towards widening the distillation bottleneck for reasoning models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul 27–Aug 1; Vienna, Austria. Volume 1: Long Papers. p. 23506–16.
11. Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. Nature. 2025;645(8081):633–8. doi:10.1038/s41586-025-09422-z.
12. Ba Z, Zhang H, Xie Z, Zuo X, Hou J. Automatic prompt engineering technology for large language models: a survey. J Front Comput Sci Technol. 2025;19(12):3131–52. (In English). doi:10.3778/j.issn.1673-9418.2502027.
13. Wang Y, Kordi Y, Mishra S, Liu A, Smith NA, Khashabi D, et al. Self-instruct: aligning language models with self-generated instructions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023 Jul 9–14; Toronto, NA, Canada. Volume 1: Long Papers. p. 13484–508.
14. Hao S, Suber M, Hu Z, Shi W, Xiong C, Griffiths TL, et al. Training large language models to reason in a continuous latent space. In: Proceedings of the International Conference on Learning Representations; 2025 Apr 24–28; Singapore.
15. Li D, Cao S, Griggs T, Liu S, Mo X, Tang E, et al. Language models can easily learn to reason from demonstrations: structure, not content, is what matters! In: Findings of the association for computational linguistics: EMNLP 2025. Kerrville, TX, USA: Association for Computational Linguistics; 2025.
16. Kang L, Zhao Z, Hsu D, Lee WS. On the empirical complexity of reasoning and planning in LLMs. In: Findings of the Association for Computational Linguistics: EMNLP 2024. Kerrville, TX, USA: Association for Computational Linguistics; 2024. p. 2843–62.
17. Duan J, Diffenderfer J, Madireddy S, Chen T, Kailkhura B, Xu K. UProp: investigating the uncertainty propagation of LLMs in multi-step agentic decision-making. arXiv:2506.17419. 2025. doi:10.48550/arxiv.2506.17419.
18. Ma CQ, Xie W, Sun WJ. Research on reinforcement learning technology: a review. Command Control Simul. 2018;40(6):68–72. (In English). doi:10.3969/j.issn.1673-3819.2018.06.015.
19. Shinn N, Cassano F, Berman E, Gopinath A, Narasimhan K, Yao S. Reflexion: language agents with verbal reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook, NY, USA: Curran Associates Inc.; 2023. p. 8634–52.

20. Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S, et al. Self-refine: iterative refinement with self-feedback. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook, NY, USA: Curran Associates Inc.; 2023. p. 46534–94.
21. Paul D, Ismayilzada M, Peyrard M, Borges L, Bosselut A, West R, et al. REFINER: reasoning feedback on intermediate representations. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics; 2024 Mar 17–22; St. Julians, Malta. Volume 1: Long Papers. p. 1100–16.
22. Chow Y, Tennenholtz G, Gur I, Zhuang V, Dai B, Thiagarajan S, et al. Inference-aware fine-tuning for best-of-N sampling in large language models. In: Proceedings of the International Conference on Learning Representations; 2025 Apr 24–28; Singapore.
23. Wu Y, Sun Z, Li S, Welleck S, Yang Y. An empirical analysis of compute-optimal inference for problem-solving with language models. In: Proceedings of the International Conference on Learning Representations; 2025 Apr 24–28; Singapore.
24. Zheng YZ, Cui X. Research on technology of integrating chain of thought in LLMs with chain of command and decision. Control Decis. 2026;41(2):432–44. (In English). doi:10.13195/j.kzyjc.2025.0534.
25. Qwen Team. Qwen3 technical report. arXiv:2505.09388. 2025.
26. Snell C, Lee J, Xu K, Kumar A. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. arXiv:2408.03314. 2024.
27. Kim Y, Park C, Jeong H, Chan YC, Xu X, McDuff D, et al. MDAgents: an adaptive collaboration of LLMs for medical decision-making. In: Proceedings of the 38th International Conference on Neural Information Processing Systems; 2024 Dec 10–15; Vancouver, BC, Canada; 2024. p. 79410–52.