

---

# PDE-AGENT: A TOOLCHAIN-AUGMENTED MULTI-AGENT FRAMEWORK FOR PDE SOLVING

---

Jianming Liu, Ren Zhu, Jian Xu, Kun Ding, Xu-Yao Zhang, Gaofeng Meng, Cheng-Lin Liu

## ABSTRACT

Solving Partial Differential Equations (PDEs) is a cornerstone of engineering and scientific research. Traditional methods for PDE solving are cumbersome, relying on manual setup and domain expertise. While Physics-Informed Neural Network (PINNs) introduced end-to-end neural network-based solutions, and frameworks like DeepXDE further enhanced automation, these approaches still depend on expert knowledge and lack full autonomy. In this work, we frame PDE solving as tool invocation via LLM-driven agents and introduce PDE-Agent, the first toolchain-augmented multi-agent collaboration framework, inheriting the reasoning capacity of LLMs and the controllability of external tools and enabling automated PDE solving from natural language descriptions. PDE-Agent leverages the strengths of multi-agent and multi-tool collaboration through two key innovations: (1) A Prog-Act framework with graph memory for multi-agent collaboration, which enables effective dynamic planning and error correction via dual-loop mechanisms (localized fixes and global revisions). (2) A Resource-Pool integrated with a tool-parameter separation mechanism for multi-tool collaboration. This centralizes the management of runtime artifacts and resolves inter-tool dependency gaps in existing frameworks. To validate and evaluate this new paradigm for PDE solving , we develop PDE-Bench, a multi-type PDE Benchmark for agent-based tool collaborative solving, and propose multi-level metrics for assessing tool coordination. Evaluations verify that PDE-Agent exhibits superior applicability and performance in complex multi-step, cross-step dependent tasks. This new paradigm of toolchain-augmented multi-agent PDE solving will further advance future developments in automated scientific computing. Our source code and dataset will be made publicly available.

## 1 Introduction

Partial Differential Equations (PDEs) are fundamental for modeling intricate phenomena in science and engineering, with applications in fluid dynamics [1], heat transfer [2], and quantum mechanics, underpinning our understanding of physical systems. Efficiently solving PDEs remains a critical bottleneck. Conventional numerical methods, finite element (FEM) [3], finite difference (FDM) [4], finite volume (FVM) [5], rely on mesh-based discretization, incur high computational costs, and face the curse of dimensionality. Physics-Informed Neural Networks (PINNs) [6, 7, 8], a promising deep learning approach, constrain neural networks to minimize PDE residuals, using automatic differentiation for precise derivative computation to avoid truncation errors and demonstrate potential for overcoming dimensional limitations. Frameworks like DeepXDE [9] have advanced their practical use. Yet current methods still require significant domain expertise and programming skills, limiting their broader scientific utility. Thus, developing automated solving frameworks to reduce entry barriers and boost efficiency is critical to advancing the field.

Large Language Models (LLMs) [10, 11] have shown strong reasoning and problem-solving abilities in general domains, with LLM-driven agents increasingly applied to disciplinary problems and automated scientific exploration [12, 13, 14], notably in text understanding [15], code generation [16, 17], and interdisciplinary QA [18]. However, their use in specialized scientific computing faces critical challenges: insufficient domain knowledge (limited coverage and poor understanding of specialized symbols/descriptions), struggles with complex computational tasks (declining performance and hallucinations), and reliance on statistical patterns in training data rather than true comprehension of physical essence, leading to unreliable results.

Analogous to humans, tool-using is key to expanding cognitive boundaries and boosting capabilities [19]. Inspired by this insight, LLM agents collaborating on tool use can bridge their capability and knowledge gaps in complex scientific computing [20]. This multi-tool collaborative approach offers a promising new paradigm for automated PDE solving: it

automatically converts high-level natural language PDE descriptions into executable task planning, tool scheduling, and response generation, boosting problem-solving automation/efficiency while enhancing solution interpretability and robustness.

Meanwhile, current research on tool-using agent [21, 22, 23] largely relies on end-to-end metrics (e.g., Pass@K, BLEU [24]) measuring task success or final output quality. While these capture overall task completion, they overly focus on static outcomes and neglect dynamic processes, failing to quantify fine-grained dimensions of tool-using ability. Notably, in scientific computing, with strong causal dependencies and high fault tolerance needs, this outcome-centric approach cannot diagnose bottlenecks or support optimization/verification of coordination strategies. Thus, there is an urgent need for multi-dimensional process evaluations that transcend end-to-end metrics to scientifically deconstruct agents' tool-using performance.

To this end, we present a multi-type PDE benchmark for tool-collaboration invocation - PDE-Bench, which contains differential equations with a total of approximately 100 test cases, along with a general multi-tiered metrics that can comprehensively assess tool invocation at the global, logical, and local-detail levels. To demonstrate the generality of our approach, we evaluate it on this benchmark and concurrently present PDE-Agent, an innovative multi-agent framework equipped with modular PDE-toolkits that we have rigorously encapsulated for automated PDEs solving. The architecture integrates task-aware decomposition, tool-oriented execution, and self-refining collaboration strategies.

PDE-Agent establishes an autonomous, closed-loop architecture from natural language problem descriptions to PDE solutions. To achieve better multi-agent collaboration, we further propose a Prog-Act, a Progressive reasoning and Acting approach. Prog-Act is a semi-dynamic planning method that enables active assistance-seeking, fixed-point verification and escalating collaboration through multi-agent design, and realizes error localization and error tracing by tool-chain constructing dynamically. In addition, we propose a resource-pool design to address the issue of implicit parameter passing in multi-tool collaboration. Fig.1 provides a task-level accuracy overview, demonstrating that our approach achieves competitive and stable performance across diverse PDE problems; detailed statistics are deferred to Section 5.

To summarize, the critical contributions of our work are the following:

1. We introduce PDE-Agent, the first LLM-driven multi-agent framework equipped with modular PDE-toolkits - rigorously encapsulated for automated PDEs solving.
2. We develop PDE-Bench, a manually curated multi-type PDE benchmark for tool-collaborative invocation, comprising differential equations across approximately 100 cases. Alongside this dataset, we propose general multi-level evaluation metrics encompassing global, logical, and local-detail-level assessments for comprehensive tool invocation evaluation.
3. We further present Prog-Act, a Progressive reasoning and Acting approach, and a resource-pool design. These innovations enhance multi-agent collaboration and multi-tool coordination, directly improving the robustness and efficiency of automated PDE solving.
4. Finally, we comprehensively evaluated and compared agent and tool invocation methods, and the experimental results demonstrate the effectiveness of PDE-Agent.

## 2 Related Work

### 2.1 Automation and Intelligence in PDE Solving

As a critical tool for describing physical systems, automated solving of PDEs remains a key challenge in scientific computing. Traditional numerical methods (e.g., FEM, FDM) [25, 26] rely on manual mesh generation [27] and discretization tuning, limiting automation for complex geometries or multi-physics problems. Deep learning has shifted focus to intelligent PDE solving, with PINNs [6] emerging as a leading framework [7, 8, 28]: by embedding PDE

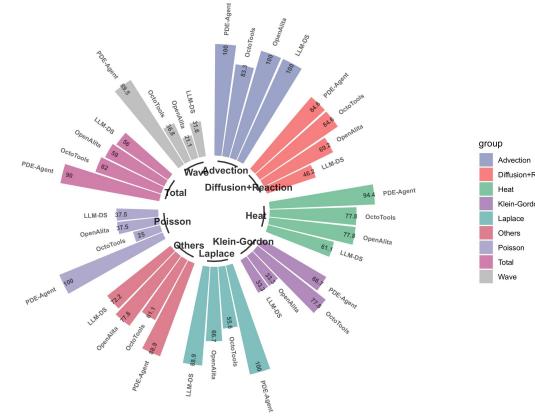


Figure 1: The competitive and stable performance across diverse PDE problems.

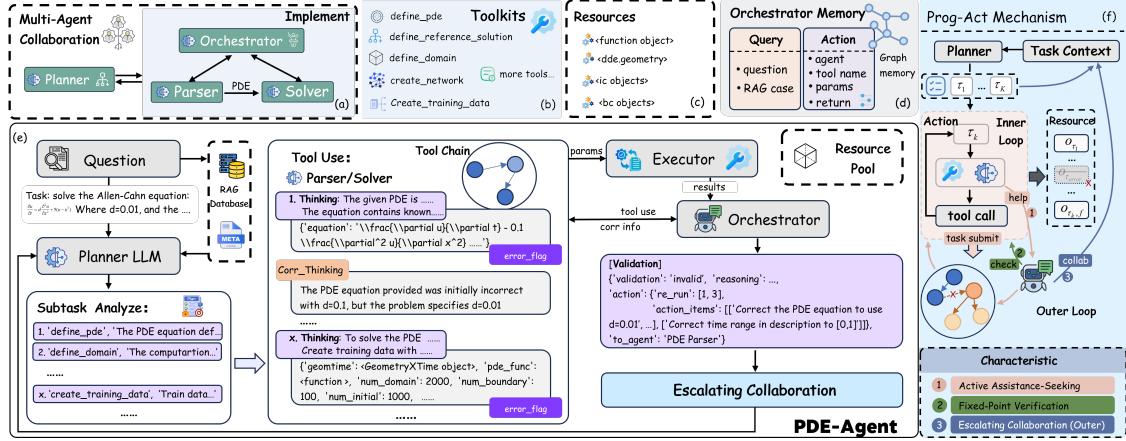


Figure 2: The architecture of PDE-Agent.

constraints into the loss function, PINNs avoid mesh-based discretization, enabling end-to-end PDE solution and marking a key advance in automation. Neural-network-based tools like NeuralPDE [29], PyDEns [30], and DeepXDE build on this, advancing automation via unified interfaces for defining PDEs, configuring networks, and optimizing training—lowering barriers for non-experts to use PINNs.

However, existing PINNs-based methods still face key limitations for fully automating PDE solving: PDEs and their constraints often require manual coding (conflicting with “fully automated” goals), and adjustments to network structures or training strategies remain heavily human-dependent, lacking systematic collaborative mechanisms. To address these, our work inherits PINNs’ core ideas and draws on DeepXDE, encapsulating PDE-toolkits to advance automation. To our knowledge, it is the first framework to automate PDE solution via coordinated multi-agent tool invocation, independent of expert intervention.

## 2.2 Tool-Augmented LLM Agents

LLMs [31, 32] have advanced significantly, driving transformative impact across scientific tasks with strong processing and reasoning capabilities. This progress has spurred LLM-driven agents [33], revolutionizing complex scientific research automation via goal-directed behavior, dynamic decision-making, and environmental interaction. However, LLMs face inherent limitations in intricate computational tasks, especially those requiring precise numerical operations, symbolic manipulations, or multi-step deductions, due to reliance on fixed parametric knowledge [34, 35], often leading to inaccuracies or “hallucinations” in high-stakes scientific computations [36]. Augmenting LLMs with external tools [37, 22] thus emerges as a promising solution, enhancing stability and interpretability by grounding outputs in rigorous algorithms rather than probabilistic predictions.

Recent advances in integrating agents with external tools have boosted automated problem-solving [38, 39]. ReAct [40] laid a foundation by unifying reasoning and acting but, focused on “immediate acting” and lacking long-term planning, limits handling complex multi-step tasks. Pre-Act [41] refined this via pre-emptive planning: generating a global multi-step plan upfront reduces inefficiencies/errors from limited perspective and boosts efficiency, particularly for cross-step dependent tasks. OctoTools [22] uses a hierarchical planning similar to Pre-Act and, as an extensible multi-agent tool framework, performs well across tasks. Yet current frameworks have limitations: (1) step-by-step validation may cause inefficiencies, redundancy, and overhead in stable tool collaboration; (2) tool dependency management focuses on explicit parameters but neglects critical implicit ones for multi-tool coordination. To address these, we propose Prog-Act and Resource-Pool to enhance multi-agent/multi-tool collaboration (details in Section 3).

## 3 Method: PDE-Agent

To enable end-to-end automated PDE solving, we propose PDE-Agent, a toolchain-augmented multi-agent framework for PDE solving. The framework aims to realize full automation of the PDE solving, encompassing problem understanding, task decomposition, tool collaboration, and dynamic validation, while integrating specifically designed mechanisms (Section 3.2 and 3.3) to optimize these types of complex, multi-step processes. Its effectiveness is validated using our PDE-Data and multi-level evaluation metrics.

### 3.1 Overall Architecture of PDE-Agent

The PDE-Agent framework comprises four core components, as illustrated in Fig.2: Planner, Parser/Solver, Executor and Orchestrator. These components are designed to operate in a closed-loop process, collectively enabling end-to-end PDE solving from natural language input.

Generally, user queries  $\mathcal{Q}$  for PDE problem are typically unstructured and often embody complex intentions. Thus, the Planner first conducts a comprehensive analysis of the PDE problem and solution requirements to generate a holistic plan, decomposing the problem into potential executable  $K$  subtasks denoted as  $\mathcal{T} = \{\tau\}_{k=1}^K$ . This decomposition ensures consistency in long-term task planning. For each subtask  $\tau = \langle k, f, g \rangle \in \mathcal{T}$ , The Planner specifies: a task ID  $k$ , a target tool  $f \in \mathcal{F}$  (where  $\mathcal{F} = \{f\}_{n=1}^N$  denotes the set of available  $N$  tools), and a subtask-specific goal/reasoning  $g$ .

Subsequently, the Parser/Solver and Executor iteratively handle each subtask, with results aggregated by the Orchestrator. To stabilize tool utilization, we adopt a tool-parameter separation mechanism: the Parser/Solver verifies tool validity, extracts task-specific parameters  $\theta_{\tau,f}$  for subtask  $\tau$  or tool  $f$ . (with the Parser extracting critical PDE objects/conditions like domain or boundary/initial conditions, and the Solver configuring neural networks to perform core solving), while the Executor invokes tool  $f$  using extracted parameters  $\theta_{\tau,f}$ , executes it, and outputs results.

After each subtask, the action  $a := \{k, f, \theta, o\}$  is synchronously submitted to the Orchestrator. As “Cognitive Elite” of PDE-Agent, the Orchestrator oversees the entire task lifecycle, with primary responsibilities including validating the accuracy of subtask executions and implementing error correction such as flagging invalid parameters or reconfiguring tools for failed subtasks. Upon completion of all subtasks, the final solution to the PDE problem is formally represented as:

$$\mathcal{S} = \text{Orchestrator}^*(\mathcal{Q}, \{\text{Executor}(\tau, f, \theta_{\tau,f})\}_{\tau \in \mathcal{T}}) \quad (1)$$

### 3.2 Multi-agent Collaboration

**Is Step-by-step Validation Necessary?** In our multi-agent architecture of PDE-Agent, ensuring coherence across multi-step task and maintaining global control over PDE solving require a holistic planning approach the Planner generates. To enhance adaptability to external feedback, existing methods adopt step-by-step validation to enable dynamic replanning; admittedly, this augments the agent’s reasoning and task adaptability. However, is step-wise verification truly necessary? Is such an overly cautious design warranted?

**Prog-Act.** Tool-augmented, LLM-driven agents simultaneously inherit the reasoning capacity of LLMs and the deterministic controllability of external tools, thereby exhibiting higher stability than the code-generation approach. Against this backdrop, step-wise validation risks introducing excessive information redundancy, leading to unwarranted consumption of temporal and computational resources. To address this trade-off, we propose the Prog-Act framework: a dual-loop error detection and correction mechanism that fully exploits multi-agent collaboration to balance adaptability and efficiency. Unlike previous step-by-step validation, Prog-Act introduces active assistance-seeking, fixed-point verification and escalating collaboration to optimize error handling while preserving task continuity, shown in Fig.2f.

The Prog-Act approach operates through two interconnected loops (dual-loop error handling):

Inner loop: localized error correction. The Orchestrator, as the central coordinator, receives action submissions, with the latest system-wide information. Rather than step-wise validating, it verifies at fixed checkpoints (e.g., after critical tool invocations or predefined subtask counts); additionally, Parser/Solver and Executor can proactively request validation via active collaboration. This establishes systematized multi-agent division of labor, aligning with human cognitive intuition. During verification, the Orchestrator analyzes current action sequence  $\mathcal{A} = \{a\}_{m=1}^M$  against the original PDE problem  $\mathcal{Q}$  for targeted error correction.

$$\text{Validate}(\mathcal{Q}, \mathcal{A}_c, \mathcal{G}) = \begin{cases} \text{Pass} & \text{if noError}(\mathcal{A}_c), \\ \text{InLoop}(\mathcal{A}_c, \mathcal{G}) & \text{otherwise} \end{cases} \quad (2)$$

where  $\mathcal{A}_c \subseteq \mathcal{A}$  is the action subset at checkpoint  $c$ , and  $\mathcal{G}$  represents the graph memory detailed in next part.

Outer loop: global plan revision. If inner-loop “collapse”, the Orchestrator formally challenges the Planner, initiating global plan revision. The Planner delivers adaptive plans through continuous refinement or, when necessary, executes full replanning. This dual-loop approach ensures efficient handling of both transient errors (via the inner loop) and systemic issues (via the outer loop).

$$\mathcal{T}_{t+1} = \text{Planner}(\mathcal{Q}, \mathcal{T}_t, \text{Feedback}(\mathcal{A}, \mathcal{G})) \quad (3)$$

where  $\text{Feedback}(\mathcal{A}, \mathcal{G})$  aggregates error and recovery actions from the inner loop.

**Graph Memory.** To enhance error traceability and cross-step coordination, PDE-Agent employs a Graph Memory structure that models tool/subtask dependencies and data flows, further augmenting the Pre-Act as the icing on the

cake. During execution, the system constructs a directed graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where nodes  $v \in \mathcal{G}$  represent tools or subtasks, and directed edges  $e \in \mathcal{E}$  represent data transmission or dependency relationships. The Orchestrator, with graph memory, can immediately identify the tool or subtask nodes tainted by erroneous data and initiates subsequent remediation upon error localization. This graph-based memory enables efficient error localization and recovery, aligning with the “progressive” nature of Prog-Act’s validation strategy.

This dual-loop, graph-augmented approach enables PDE-Agent to balance efficiency and robustness, addressing the limitations of both overly cautious step-by-step validation and purely end-to-end verification. By leveraging multi-agent collaboration and structured memory, Prog-Act ensures coherent execution of complex, multi-step, and tool/subtask-dependent process in PDE solving.

### 3.3 Multi-tool Coordination

Effective coordination of specialized tools is critical for complex multi-step tasks like PDE solving. However, existing tool-augmented methods predominantly rely on explicit parameter passing, extracting downstream tool parameters from problem descriptions or prior outputs and encoding them directly into commands, failing to handle implicit dependencies, such as runtime artifacts (e.g., function objects, class instances) from preceding tools that cannot be serialized, a key limitation in multi-tool collaboration unmanaged by current frameworks.

To bridge this gap, we introduce the Resource-Pool ( $\mathcal{R}$ ), which originally serves as a repository for “creations” generated by tools integrated with a tool-parameter separation. Remarkably, it shows strong extensibility, supporting flexible integration of any resources required by the system. As such, this component can also serve as a dynamic repository for shared resources, accessible to all components within the framework, and allowing the Parser/Solver and the Executor to access and update interdependent resources. Formally, with the tool’s results  $o_{\tau,f}$ , the Resource-Pool is updated by:

$$\mathcal{R} \leftarrow \mathcal{R} \cup \{o_{\tau,f}\} \quad (4)$$

By centralizing implicit resources management, the Resource-Pool enables seamless multi-tool coordination, aligning naturally with the tool-parameter separation architecture to address a critical limitation of existing frameworks.

## 4 Benchmark and Evaluation Metrics

In this section, we introduce our PDE dataset (PDE-Data) and details of the multi-level evaluation metrics we propose to assess multi-agent and multi-tool coordination effectiveness in automated problem solving.

### 4.1 PDE-Data

**Design Motivation.** Existing PDE datasets (e.g., PDEBench [42]) primarily serve the training and validation of numerical solving methods, concentrating on the “equation formulation - solution data” mapping. However, they lack support for the full workflow of “natural language problem description  $\rightarrow$  tool-collaborative solving”. To address this gap and the evaluation requirements of PDE-Agent, we constructed PDE-Data, aiming to provide a test dataset covering diverse PDE problems with full workflow annotations. It is manually curated, with detailed examples provided in the Appendix, enabling comprehensive evaluation of entire pipeline: task completion  $\rightarrow$  tool collaboration  $\rightarrow$  tool parameters and returns.

**Data composition.** PDE-Data includes 100 differential equation test cases, with detailed distribution in Table 1. Distributed to balance fundamental representativeness and scenario diversity, its “foundational cases as mainstay, specialized scenarios as supplement” composition supports assessing agents’ core PDE-solving capabilities and generalization in complex/niche scenarios, providing comprehensive multi-dimensional evaluation data.

### 4.2 Evaluation Metrics

We propose three categories of evaluation metrics to comprehensively assess PDE-Agent: global task completion (a global metric measured via Pass@k or task success rate); local tool invocation assessment (a detailed evaluation of tool invocation specifics, including Semantic Textual Similarity (STS) and BERTScore for tool parameters and outputs); and logical collaboration process assessment. Detailed descriptions of the first two are provided in the Appendix. Here, we emphasize the third category: the newly proposed logical collaboration process assessment, a process-oriented evaluation that gauges the rationality of tool interaction logic.

**Logical Collaboration Process Assessment.** Evaluating logical rationality of tool invocations, we focus on coherence of tool interactions, formalizing collaboration as a directed graph (tool chain)  $\mathcal{G}$  with nodes  $\mathcal{V}$  (tools) and directed

edges  $\mathcal{E}$  (inter-tool data flow). Let  $\mathcal{G}_p(\mathcal{V}_p, \mathcal{E}_p)$  be the ground-truth pattern graph (from PDE-Data) and  $\mathcal{G}_e(\mathcal{V}_e, \mathcal{E}_e)$  be PDE-Agent's actual execution graph. Assessment has two core components: (1) Tool chain completeness, scored via (partial) subgraph isomorphism and PageRank to preserve critical tools/data flows; (2) Tool chain error rate, quantified via graph edit distance and similarity to measure deviations from expected logic.

**(Partial) Subgraph Isomorphism-Based Evaluation.** This metric combines node-edge matching and partial subgraph ( $\mathcal{G}_{sub}$ ) structural similarity to assess logical integrity:

Node and edge matching: Evaluates the completeness of tools and data flows. It quantifies the proportion of nodes (tools) and edges (data dependencies) in  $\mathcal{G}_e$  that match their counterparts in  $\mathcal{G}_p$

$$R_{\mathcal{V}} = \frac{|\mathcal{V}_e \cap \mathcal{V}_p|}{|\mathcal{V}_p|}, \quad R_{\mathcal{E}} = \frac{|\mathcal{E}_e \cap \mathcal{E}_p|}{|\mathcal{E}_p|} \quad (5)$$

**Structural Similarity of Partial Subgraphs:** Assesses the logical coherence of internal toolchains through connectivity metrics, role node similarity, critical path similarity, and execution order similarity. Additional details on these metrics and other evaluation criteria are provided in the Appendix.

**PageRank-JS similarity metric.** The proposed metric quantifies topological congruence by comparing PageRank [43] distributions via Jensen-Shannon divergence. Firstly, we get the PageRank vector  $\mathbf{p}_k$ , which captures the relative importance of each tool node. Then, we compute the Jensen-Shannon divergence between distributions.

$$D_{JS}(\mathbf{p}_{\mathcal{V}_p} \| \mathbf{p}_{\mathcal{V}_e}) = \frac{1}{2} \left[ D_{KL}(\mathbf{p}_{\mathcal{V}_p}^{\text{norm}} \| M) + D_{KL}(\mathbf{p}_{\mathcal{V}_e}^{\text{norm}} \| M) \right] \quad (6)$$

where  $M = \frac{1}{2}(\mathbf{p}_{\mathcal{V}_p}^{\text{norm}} + \mathbf{p}_{\mathcal{V}_e}^{\text{norm}})$ ,  $\mathbf{p}_k^{\text{norm}} = \frac{\mathbf{p}_k}{\|\mathbf{p}_k\|_1}$  is the L1-normalize vector and  $D_{KL}$  is the Kullback-Leibler divergence.

Finally, we can obtain the similarity metric

$$\text{Sim}(\mathcal{G}_p, \mathcal{G}_e) = 1 - \sqrt{D_{JS}(\mathbf{p}_{\mathcal{V}_p} \| \mathbf{p}_{\mathcal{V}_e})} \quad (7)$$

**Graph Edit Distance-Based Similarity.** We propose a systematic similarity metric based on Normalized Graph Edit Distance (NGED) to quantify similarity between tool-chain. We calculated the maximum possible edit distance  $\text{Base}(\mathcal{G}_e, \mathcal{G}_p)$  and can get the NGED

$$\begin{aligned} \text{Base}(\mathcal{G}_e, \mathcal{G}_p) &= |\mathcal{V}_e| + |\mathcal{E}_e| + |\mathcal{V}_p| + |\mathcal{E}_p| \\ \text{NGED}(\mathcal{G}_e, \mathcal{G}_p) &= \min \left( \frac{\text{GED}(G_1, G_2)}{\text{Base}(G_1, G_2)}, 1 \right) \end{aligned} \quad (8)$$

Then we use the exponential mapping to transform NGED to the similarity score

$$S(\mathcal{G}_e, \mathcal{G}_p) = 1 - \frac{1 - e^{-\alpha \cdot \text{NGED}(\mathcal{G}_e, \mathcal{G}_p)}}{1 - e^{-\alpha}} \quad (9)$$

where  $\alpha > 0$  is a sensitivity parameter controlling the steepness of the similarity curve and we use  $\alpha = 2$  in our experiments.

**Graph embedding similarity based evaluation.** The Graph Embedding Similarity employs the Node2Vec algorithm [44] to generate latent representations of tool nodes, and compute a graph-level embedding  $\mathbf{g}_G$  through centrality-weighted aggregation of node embeddings.

Then we compute the raw similarity using cosine similarity and transform the raw similarity into a normalized score  $S \in [0, 1]$  using a parameterized sigmoid transformation

$$\begin{aligned} \text{sim}_{\text{raw}}(\mathcal{G}_e, \mathcal{G}_p) &= \frac{\mathbf{g}_{\mathcal{G}_e} \cdot \mathbf{g}_{\mathcal{G}_p}}{\|\mathbf{g}_{\mathcal{G}_e}\| \cdot \|\mathbf{g}_{\mathcal{G}_p}\|} \\ S(G_A, G_B) &= \frac{\sigma(k \cdot \text{sim}_{\text{raw}}) - \sigma(-k)}{\sigma(k) - \sigma(-k)} \end{aligned} \quad (10)$$

where  $\sigma(x) = \frac{1}{1+e^{-x}}$  represents sigmoid function and  $k$  is slope parameter controlling score distribution ( $k = 1$  by default).

Category	Total	PDE-A	Octo	O-Alita	LLM-DS
Heat Equation	18	17	14	14	11
Diffusion	6	4	5	2	2
Diffusion-Reaction	7	7	6	7	4
Wave Equation	19	17	7	4	6
Klein-Gordon	9	6	7	3	3
Advection	6	6	5	6	6
Burgers	4	3	4	3	3
Laplace	9	9	5	6	8
Poisson	8	8	2	3	3
Allen-Cahn	6	6	4	6	5
Others	8	7	3	5	5
Overall	100	90%	62%	59%	56%

Table 1: Overall performance by PDE type and methods.

## 5 Experiments

To validate the effectiveness of PDE-Agent in automatic PDE solving from natural language, we conduct comprehensive experiments on our PDE-Data. The experiments aim to : (1) Compare overall performance with methods from diverse perspectives; (2) Analyze performance through multi-dimensional metrics; (3) Verify the contributions of mechanism via ablation studies and case analyses. Note that detailed sample-wise statistics and analyses from the experiments are provided in the Appendix.

### 5.1 Overall Performance Comparison

**Experimental Setup.** We use PDE-Data, a curated dataset, balancing commonly studied PDE categories and less frequent types for robust evaluation of the framework’s ability to handle diverse PDE scenarios. We compare PDE-Agent against three

- **OctoTools:** A recent open-source agent-toolbox framework designed to streamline multi-tool workflows in complex reasoning tasks, which has achieved strong performance across various complex tasks.
- **OpenAlita:** An agent system characterized by minimal predefinition and maximal self-evolution, which attempts to construct functions via web-search (similar to code generation) for agent tool augmentation.
- **LLM-DS:** A pure LLM baseline (DeepSeek-v3) without explicit tool augmentation, relying solely on internal reasoning for PDE solving.

For all compared methods, we adopt the same LLM (DeepSeek-v3) and hyperparameters to ensure fairness. We adopt global evaluation metrics (primarily Pass@1) for overall performance, and conduct in-depth analyses of tool-augmented methods using our multi-level metrics. Notably, we adapted OctoTools and OpenAlita with the Resource-Pool; without this adaptation, they would be unable to complete toolchain collaboration tasks.

Table 1 presents the Pass@1 scores of PDE-Agent and baselines across PDE-Data. PDE-Agent achieves an overall success rate of 90%, outperforming other methods by significant margins. There are some key observations:

- The pure LLM baseline (LLM-DS) performs poorly due to its inability to leverage tools - critical for reliable symbolic abstraction (e.g., define\_pde) and complex computations, both of which are essential for PDE solving.
- OpenAlita’s automatic tool construction functionality currently relies heavily on code-generation capabilities, with limited abstraction of tool functionalities. This hinders its performance beyond basic tool setup.
- OctoTools, which employs a general step-by-step validation framework without toolchain augmentation, struggles with two key limitations: (1) Failure to trace dependencies across long-range, cross-step tasks; (2) Information redundancy from step-wise validation, misinterpretation of task intent and premature task termination.
- PDE-Agent’s superiority stems from two core mechanisms: the Prog-Act framework (enabling multi-agent collaboration to handle unexpected errors during solving) and the Resource-Pool (facilitating tool coordination through effective management of implicit parameters).

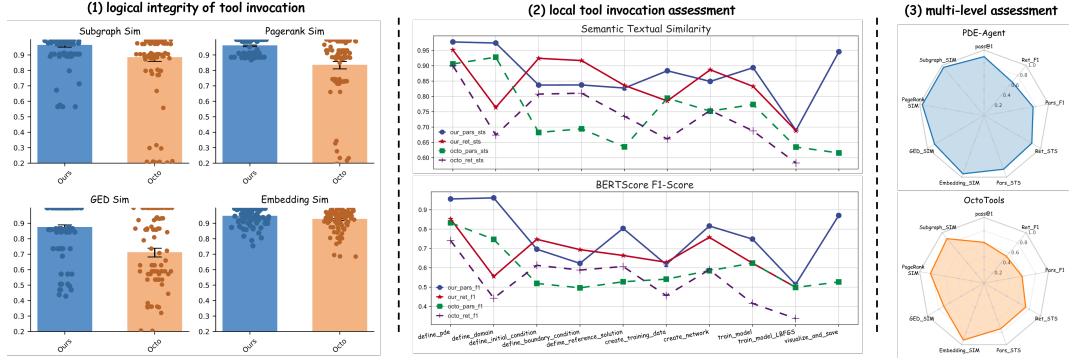


Figure 3: The multi-level evaluation.

Category	Prog-Act			diff-LLM		
	Full	w/o	step	DS	o4-mini	Gemini
Heat Equation (18)	17	15	16	17	14	15
Diffusion (6)	4	2	6	4	5	4
Diffusion-Reaction (7)	7	5	6	7	7	6
Wave Equation (19)	17	12	15	17	16	18
Klein-Gordon (9)	6	7	6	6	6	7
Advection (6)	6	4	6	6	4	6
Burgers (4)	3	3	2	3	3	3
Laplace (9)	9	9	9	9	9	7
Poisson (8)	8	5	6	8	4	5
Allen-Cahn (6)	6	6	6	6	6	6
Others (8)	7	8	8	7	7	8
Overall	90%	76%	86%	90%	81%	85%

Table 2: Ablation study on the Prog-Act and different LLM for PDE-Agent.

## 5.2 In-Depth Comparison with OctoTools

OctoTools is the most comparable baseline, as it focuses on tool-augmented, multi-step reasoning. However, overall success rates alone are insufficient to evaluate performance in multi-step tasks, as they fail to capture task completion quality, tool invocation logic, or step-by-step tool utilization. We thus conduct a fine-grained analysis across multiple distinct dimensions (Fig.3).

**Logical collaboration process assessment.** We evaluate the logical integrity of tool invocation performance using graph-based metrics which is detailed in Section 4.2 which assess logical completeness and similarity. The results of these metrics across all test cases (Fig.3(1)) demonstrate that PDE-Agent outperforms OctoTools in all four dimensions. Notably, OctoTools exhibits moderate task completion quality despite lower overall success rates—an insight not captured by the global Pass@1 metric. More detailed results are provided in the Appendix.

**Local tool invocation assessment.** We measure Semantic Textual Similarity (STS) and BERTScore of the parameters and outputs for tool invocation to evaluate the accuracy of step-by-step tool utilization. STS can reflect alignment with task requirements. BERTScore, including precision, recall and F1-score, quantifies the semantic alignment. As shown in Fig.3(2), PDE-Agent achieves higher semantic similarity and BERTScore in both parameters and outputs, indicating more stable tool invocation and better mastery of tool functionalities. This confirms the effectiveness of its mechanisms in ensuring accurate, context-aware tool utilization. Furthermore, this assessment can serve two purposes: on one hand, it functions as an evaluation of the entire system; on the other hand, it enables researchers to conduct targeted optimization of tools or enhance agents' capabilities in the specific aspect. Detailed evaluations and analyses are provided in Appendix.

The Fig.3(3) summarizes multi-level evaluation metrics to reflect the overall performance of the methods. These results collectively demonstrate the PDE-Agent not only achieves higher overall success rates but also exhibits superior logical coherence in tool collaboration and more accurate local tool invocation.

### 5.3 Ablation Studies and Case Analysis

The Resource-Pool design enables the propagation of runtime artifacts (e.g., function objects, class instances) between tools in collaborative workflows. In complex multi-tool tasks, effective information flow cannot be achieved without such centralized management. Thus, the Resource-Pool is indispensable for coherent tool collaboration.

To further clarify the role of the Prog-Act in PDE-Agent, we conducted ablation studies by disabling Prog-Act and comparing it with a step-wise dynamic validation strategy, contrasting it with the fully-equipped framework.

**Effectiveness of Prog-Act.** Table 2 presents the performance of three PDE-Agent variants: the complete framework, PDE-Agent without Prog-Act (w/o Prog-Act), and a variant using step-wise validation (replacing Prog-Act’s dual-loop mechanism).

Removing Prog-Act reduces the success rate by 14%. The framework reverts to a static tool chain, becoming unstable and unable to recover from errors (e.g., invalid parameter propagation). Compared to the step-wise validation variant, our framework achieves comparable performance while consuming significantly fewer resources. This confirms that Prog-Act’s dual-loop strategy (localized correction for transient errors and global revision for systemic issues) avoids the redundancy of step-wise validation, balancing efficiency and robustness.

**Generalizability Across Different LLMs.** We also tested its performance across different underlying LLMs to further validate the generalizability of PDE-Agent. Specifically, we evaluated PDE-Agent using three representative models: DeepSeek-v3 [45] (our primary test model), GPT-4o-mini [46] and Gemini [47], maintaining consistent hyperparameters and tool configurations across all tests. Table 2 presents the overall success rates across these LLMs.

**Case Analysis.** And we show the Prog-Act mechanism, encompassing active collaboration in the inner loop, global plan revision in the outer loop, alongside dynamic toolchain construction and error tracing.

## 6 Conclusion

In this work, we introduce PDE-Agent, a novel toolchain-augmented multi-agent collaboration framework for fully automated PDE solving, incorporating a Prog-Act framework with graph memory to enhance multi-agent collaboration and a Resource-Pool with tool-parameter separation to improve multi-tool coordination. We also developed PDE-Data, a multi-type benchmark with diverse PDE problems and annotations, and a multi-level metrics. These contributions advance automated PDE solving by providing a robust framework tailored to toolchain-augmented multi-agent challenges and standardized evaluation infrastructure, with experimental results validating PDE-Agent’s superiority over existing methods in overall success and fine-grained tool utilization.

## References

- [1] James P Howard et al. Data-driven modeling & scientific computation: methods for complex systems & big data. *Journal of Statistical Software*, 67:1–3, 2015.
- [2] Navid Zobeiry and Keith D Humfeld. A physics-informed machine learning approach for solving heat transfer equation in advanced manufacturing and engineering applications. *Engineering Applications of Artificial Intelligence*, 101:104232, 2021.
- [3] Gennadiy P Nikishkov. Introduction to the finite element method. *University of Aizu*, pages 1–70, 2004.
- [4] DM Causon and CG Mingham. *Introductory finite difference methods for PDEs*. Bookboon, 2010.
- [5] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite volume methods. *Handbook of numerical analysis*, 7:713–1018, 2000.
- [6] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.
- [7] Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics-informed neural networks: Where we are and what’s next. *Journal of Scientific Computing*, 92(3):88, 2022.
- [8] Edgar Torres and Mathias Niepert. Survey: Adaptive physics-informed neural networks. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024.
- [9] Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM review*, 63(1):208–228, 2021.
- [10] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- [12] Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025.
- [13] Samuel Schmidgall and Michael Moor. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*, 2025.
- [14] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*, 2023.
- [15] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 2021.
- [16] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [17] Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. *arXiv preprint arXiv:2401.07339*, 2024.
- [18] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.
- [19] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*, 2024.
- [20] Beichen Zhang, Kun Zhou, Xilin Wei, Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. Evaluating and improving tool-augmented computation-intensive math reasoning. *Advances in Neural Information Processing Systems*, 36:23570–23589, 2023.
- [21] Yuxuan Chen, Long Zhang, Xu Zhu, Hua Zhou, and Zhuyin Ren. Optmetaopenfoam: Large language model driven chain of thought for sensitivity analysis and parameter optimization based on cfd. *arXiv preprint arXiv:2503.01273*, 2025.

- [22] Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. Octotools: An agentic framework with extensible tools for complex reasoning. *arXiv preprint arXiv:2502.11271*, 2025.
- [23] Jiahao Qiu, Xuan Qi, Tongcheng Zhang, Xinzhe Juan, Jiacheng Guo, Yifu Lu, Yimin Wang, Zixin Yao, Qihan Ren, Xun Jiang, et al. Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution. *arXiv preprint arXiv:2505.20286*, 2025.
- [24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [25] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method: Its Basis and Fundamentals*. The Finite Element Method. Butterworth-Heinemann, 2013.
- [26] Randall J LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.
- [27] Dimitri J. Mavriplis. Chapter 7 - mesh generation and adaptivity for complex geometries and flows. In Roger Peyret, editor, *Handbook of Computational Fluid Mechanics*, pages 417–459. Academic Press, London, 1996.
- [28] Hao Zhongkai, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, Yichi Zhang, Songming Liu, Lu Lu, et al. Pinnacle: A comprehensive benchmark of physics-informed neural networks for solving pdes. *Advances in Neural Information Processing Systems*, 37:76721–76774, 2024.
- [29] Christopher Rackauckas and Qing Nie. Differentialequations. jl—a performant and feature-rich ecosystem for solving differential equations in julia. *Journal of Open Research Software*, 5(1):15–15, 2017.
- [30] Alexander Koryagin, Roman Khudorozkov, and Sergey Tsimfer. Pydens: A python framework for solving differential equations with neural networks. *arXiv preprint arXiv:1909.11544*, 2019.
- [31] Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*, 57(6):1–38, 2025.
- [32] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.
- [33] Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.
- [34] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2022.
- [35] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343, 2025.
- [36] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143, 2023.
- [37] Zhuocheng Shen. Llm with tools: A survey. *arXiv preprint arXiv:2409.18807*, 2024.
- [38] Shen Gao, Zhengliang Shi, Minghang Zhu, Bowen Fang, Xin Xin, Pengjie Ren, Zhumin Chen, Jun Ma, and Zhaochun Ren. Confucius: Iterative tool learning from introspection feedback by easy-to-difficult curriculum. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18030–18038, 2024.
- [39] Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. What are tools anyway? a survey from the language model perspective. In *First Conference on Language Modeling*, 2024.
- [40] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
- [41] Mrinal Rawat, Ambuje Gupta, Rushil Goomer, Alessandro Di Bari, Neha Gupta, and Roberto Pieraccini. Pre-act: Multi-step planning and reasoning improves acting in llm agents. *arXiv preprint arXiv:2505.09970*, 2025.
- [42] Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Matthias Niepert. Pdebench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022.
- [43] Fan Chung. A brief survey of pagerank algorithms. *IEEE Trans. Netw. Sci. Eng.*, 1(1):38–42, 2014.

- [44] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.
- [45] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [46] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [47] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.