SALCA-IB: Self-Adaptive LLM-Driven Continuous Learning Agent for IB Network Failure Prediction

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—InfiniBand Network (IB Network) failure prediction is crucial in high-performance computing and data center operations, yet faces significant challenges due to environmental complexity and dynamicity, such as scarcity of failure data and susceptibility of network feature distributions to external factors. This paper introduces SALCA-IB (Self-Adaptive LLM-Driven Continuous Learning Agent for IB Network Failure Prediction), an innovative adaptive failure prediction agentic system. SALCA-IB utilizes a Large Language Model (LLM) as its planning core, combined with traditional machine learning methods, to achieve autonomous prediction and optimization. The system's main innovations include: (1) LLM-driven autonomous data selection and model optimization; (2) A fusion memory system integrating short-term and long-term memory; and (3) LLM-supported automatic evaluation feedback and closed-loop optimization. Experimental results show that compared to traditional methods, SALCA-IB improves prediction accuracy by X% and demonstrates a Y-fold increase when facing changes in network feature distributions. Our code is available at XXXX.github.com.**

***Index Terms*—IB Network, Large Language Model, Autonomous Agent, Memory System**

I. Introduction

High-performance computing and modern data centers heavily rely on InfiniBand (IB) networks for their superior performance in low-latency, high-bandwidth communication. As the backbone of these critical infrastructures, IB networks' reliability directly impacts the overall system performance and service availability. However, network failures can lead to severe service disruptions and significant performance degradation, making failure prediction increasingly crucial for maintaining system reliability and operational efficiency.

Despite its importance, IB network failure prediction faces several significant challenges. First, failure data in IB networks is inherently scarce, as failures are relatively rare events, making it difficult to build robust prediction models using traditional machine learning approaches. Second, network feature distributions are highly dynamic and susceptible to various external factors, such as environmental conditions, hardware aging, and maintenance activities. Third, existing prediction systems often lack the ability to adapt to these changing conditions, resulting in degraded performance over time.

Traditional approaches to network failure prediction primarily rely on static machine learning models or rule-based systems (ADD REF). While these methods have shown some success in controlled environments, they struggle to maintain performance in real-world scenarios where network characteristics evolve continuously (ADD REF). Moreover, existing solutions often operate as black boxes, providing limited interpretability and failing to leverage historical experience effectively for continuous improvement.

The emergence of Large Language Models (LLMs) presents new opportunities for addressing these challenges. LLMs have demonstrated remarkable capabilities in complex reasoning and planning tasks (ADD REF) , suggesting their potential for orchestrating adaptive prediction systems. Additionally, recent advances in memory systems and continuous learning architectures (ADD REF) have shown promise in handling dynamic environments, though their application to network failure prediction remains largely unexplored.

To address these challenges, we propose SALCA-IB (Self-Adaptive LLM-Driven Continuous Learning Agent for IB Network Failure Prediction), an innovative system that combines the reasoning and planning capabilities of LLMs with traditional machine learning models in a unified, adaptive framework. SALCA-IB introduces several key innovations that directly address the aforementioned challenges: (1) To

Identify applicable funding agency here. If none, delete this.

tackle the data scarcity issue, we develop an LLM-driven planning core that intelligently selects and utilizes limited training data, while orchestrating multiple lightweight models to maximize the value of available data; (2) To handle dynamic feature distributions, we design a dual-memory system that integrates both short-term and long-term experiences, enabling the system to capture and adapt to evolving network characteristics while maintaining historical knowledge; (3) To overcome the limitations of static prediction systems, we implement a continuous learning mechanism that enables real-time model updates and performance optimization, ensuring sustained prediction accuracy even as network conditions change.

Experimental results demonstrate that SALCA-IB outperforms traditional methods in terms of prediction accuracy and adaptability, achieving X% higher accuracy and Y-fold improvement in prediction performance when facing network feature distribution changes. Ablation studies further confirm the significant contributions of both the LLM-driven planning and dual-memory system components.

To conclude, the main contributions of this paper are threefold:

- We propose an innovative LLM-driven agent architecture (SALCA-IB) for IB network failure prediction. The system uniquely leverages LLM as a high-level planning core to orchestrate model selection, parameter optimization, and continuous learning strategies, while employing traditional machine learning models as efficient executors for real-time prediction tasks.
- We design a novel dual-memory fusion system that seamlessly integrates short-term and long-term memory mechanisms. This sophisticated memory architecture enables rapid adaptation to dynamic network changes while preserving and leveraging valuable historical knowledge, significantly enhancing the system's robustness and adaptability.
- We conduct comprehensive experiments on real-world IB network datasets to rigorously validate SALCA-IB's effectiveness. Through extensive ablation studies and comparative analyses, we demonstrate the substantial contributions of both the LLM-driven framework and the dual-memory system components to the overall system performance.

II. Related Work

A. IB Network Failure Prediction

Network failure prediction, particularly in IB networks, has been extensively studied due to its critical importance in maintaining system reliability. Traditional approaches primarily rely on statistical methods and machine learning models. XXX proposed a statistical analysis framework for predicting network failures based on historical performance metrics. XXX developed a deep learning approach using LSTM networks to capture temporal dependencies in network behavior patterns. However, these methods often struggle with the inherent data scarcity in failure scenarios and lack adaptability to changing network conditions.

More recent work has attempted to address these challenges through ensemble methods and transfer learning. XXX introduced a multi-model ensemble approach to improve prediction robustness under limited data conditions. XXX explored transfer learning techniques to leverage knowledge from similar network environments. Despite these advances, existing methods still face significant challenges in handling dynamic network environments and maintaining long-term prediction accuracy.

B. LLM in System Planning and Optimization

The application of Large Language Models (LLMs) in system planning and optimization represents an emerging research direction. XXX demonstrated LLM's capability in generating optimization strategies for complex systems, while XXX explored using LLMs for automated system configuration and parameter tuning. These studies highlight LLM's potential in understanding system behaviors and generating sophisticated planning strategies.

In the context of network systems, XXX pioneered the use of LLMs for network management and optimization. XXX further showed how LLMs can be effectively combined with traditional machine learning models to enhance system performance. However, the application of LLMs specifically for network failure prediction remains largely unexplored, particularly in terms of continuous learning and adaptation.

C. Memory-Augmented Learning Systems

Memory mechanisms have proven crucial for enhancing learning systems' long-term performance and adaptability. XXX introduced a dual-memory architecture that separates short-term and long-term memory components, enabling both rapid adaptation and stable long-term learning. XXX developed a memory-augmented neural network that demonstrates superior performance in dynamic environments.

Recent advances in memory systems have focused on efficient memory retrieval and utilization. XXX proposed an attention-based memory access mechanism that improves memory utilization efficiency. XXX developed a hierarchical memory structure that enables more effective knowledge transfer across different tasks. These works provide valuable insights for designing memory systems, though their application in network failure prediction contexts remains limited.

D. Continuous Learning for Network Systems

Continuous learning in network systems presents unique challenges due to evolving network conditions and changing failure patterns. XXX proposed an online learning framework that continuously updates prediction models based on new observations. XXX developed an adaptive learning system that automatically adjusts to network feature distribution changes.

Recent work has increasingly focused on handling concept drift in network environments. XXX introduced a drift detection mechanism that triggers model updates when significant changes are detected. XXX proposed a sliding window approach that maintains prediction accuracy under varying

network conditions. While these methods show promise, they often lack the sophisticated planning capabilities needed for complex network environments.

Unlike previous work, our SALCA-IB system uniquely combines LLM-driven planning with a dual-memory architecture, enabling both intelligent strategy generation and effective knowledge retention. This novel approach addresses the limitations of existing methods while providing superior adaptability and prediction accuracy in dynamic IB network environments.

Algorithm 1 SALCA-IB

Require: IB network data $\mathcal{D}$, LLM $\mathcal{L}$, Model set $\mathcal{M}$, Short-term memory Mem_{short}, Long-term memory Mem_{long}, Max iterations T_{max}, Performance improvement threshold δ

$D_{train} = \mathcal{LLM}(\mathcal{D}, Mem_{long})$ ▷ Train data selection
$(M, \theta, S) = \mathcal{LLM}(\mathcal{M}, Mem_{long})$ ▷ Model, hyper parameters, and integration strategy selection
$M = f_{train}(M, \theta, S, D_{train})$ ▷ Model training
$perf_{prev} = f_{eval}(M, D_{val})$ ▷ Initial performance evaluation
$t = 0$
while $t < T_{max}$ **and** $\Delta_{perf} > \delta$ **do**
  $P = M(D_{new})$ ▷ Failure prediction
  $E = f_{eval}(P, D_{actual})$ ▷ Evaluation
  $F = \mathcal{LLM}(E, M, \theta, S)$ ▷ Generate feedback
  $Mem_{short} = f_{update}(Mem_{short}, P, E, F)$ ▷ Memory Update
  $Mem_{long} = f_{update}(Mem_{long}, F, M, \theta, S)$
  $(M, \theta, S, D_{train}) = \mathcal{LLM}(Mem_{short}, Mem_{long}$ ▷ Adaptive optimization
  $M = f_{online\_train}(M, \theta, S, D_{new})$ ▷ Online model update
  $perf_{current} = f_{eval}(M, D_{val})$ ▷ Current performance evaluation
  $\Delta_{perf} = perf_{current} - perf_{prev}$
  $perf_{prev} = perf_{current}$
  $t = t + 1$
end while
return $M, \theta, S, Mem_{short}, Mem_{long}$ ▷ Return evolved Agent

III. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

IV. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections IV-A–IV-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the {eqnarray} equation environment. Use {align} or {IEEEeqnarray} instead. The {eqnarray} environment leaves unsightly spaces around relation symbols.

Please note that the {subequations} environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a \label command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a \label command should not go before the caption of a figure or a table.

Do not use \nonumber inside the {array} environment. It will not stop equation numbers inside {array} (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when

Fig. 1. Example of a figure caption.

writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.