HPC-GPT: Integrating Large Language Model for High-Performance Computing

Xianzhong Ding
xding5@ucmerced.edu
University of California, Merced
Merced, CA, USA
Argonne National Laboratory
Lemont, IL, USA

Le Chen
lechen@iastate.edu
Iowa State University
Ames, IA, USA
Lawrence Livermore National Laboratory
Livermore, CA, USA

Murali Emani
memani@anl.gov
Argonne National Laboratory
Lemont, IL, USA

Chunhua Liao, Pei-Hung Lin,
Tristan Vanderbruggen
{liao6,lin32,vanderbrugge1}@llnl.gov
Lawrence Livermore National Laboratory
Livermore, CA, USA

Zhen Xie
zxie3@binghamton.edu
Binghamton University
Binghamton, NY, USA

Alberto E. Cerpa, Wan Du
{acerpa,wdu3}@ucmerced.edu
University of California, Merced
Merced, CA, USA

ABSTRACT
Large Language Models (LLMs), including the LLaMA model, have exhibited their efficacy across various general-domain natural language processing (NLP) tasks. However, their performance in high-performance computing (HPC) domain tasks has been less than optimal due to the specialized expertise required to interpret the model’s responses. In response to this challenge, we propose HPC-GPT, a novel LLaMA-based model that has been supervised fine-tuning using generated QA (Question-Answer) instances for the HPC domain. To evaluate its effectiveness, we concentrate on two HPC tasks: managing AI models and datasets for HPC, and data race detection. By employing HPC-GPT, we demonstrate comparable performance with existing methods on both tasks, exemplifying its excellence in HPC-related scenarios. Our experiments on open-source benchmarks yield extensive results, underscoring HPC-GPT’s potential to bridge the performance gap between LLMs and HPC-specific tasks. With HPC-GPT, we aim to pave the way for LLMs to excel in HPC domains, simplifying the utilization of language models in complex computing applications.

CCS CONCEPTS
• Software and its engineering; • Computing methodologies → Artificial intelligence; Parallel algorithms;

KEYWORDS
High-performance Computing, Data Race Detection, Large Language Model, OpenMP, Neural Network.

1 INTRODUCTION
Deep learning has found application across various domains and has been utilized in a multitude of applications, including code generation [30, 32], building control [21, 22], irrigation scheduling [19, 20], and even in tasks related to epistemic uncertainty and occupancy estimation [6, 41]. These applications showcase the versatility and adaptability of deep learning techniques. On the other hand, language models (LMs), particularly large language models (LLMs) trained on extensive textual data, have recently demonstrated remarkable capabilities in various natural language processing and visualization tasks, reflecting their growing importance in the field of artificial intelligence. They have also been widely used to process programming languages due to the similarities between natural languages and programming languages. Based on an LLM trained on code [16], GitHub provides an AI assistant for developing software.

With the surging popularity of LLMs, the high-performance computing (HPC) community is exploring their potential to tackle various HPC challenges like programming language processing, parallel programming, and question answering. However, existing pre-trained LLMs were initially designed for general tasks, such as dialogue and article summarization, which limits their effectiveness in HPC applications due to a lack of relevant domain knowledge. Several attempts have been made to employ LLMs for HPC tasks. For instance, Godoy et al. [24] evaluated the capacity of OpenAI Codex [4] via Copilot for generating HPC numerical kernels, while Chen et al. [12] developed pipelines to support common HPC tasks like code similarity analysis and parallelism detection. However, they are still in the process of investigating the adequacy of existing LLMs
for HPC tasks. The general LLMs perform poorly on specific HPC tasks. For example, in question-answering tasks related to HPC, such as OpenMP Q&A [12], even the latest ChatGPT fails to provide accurate answers. This is because general LLMs lack the specific knowledge required to solve HPC-related challenges, despite being trained on large datasets that may include HPC-related information. Although the generalization ability of LLMs improves with more diverse training data from various domains, their performance in a specific domain, like HPC, reduces.

In this paper, we aim to explore and train an open-source LLM tailored for HPC applications. Our focus revolves around two general applications, each comprising two subtasks. The first application focuses on managing AI models and datasets for programming language processing (PLP) [23] and MLPerf [42] tasks. Through datasets and model information collection, we aim to help users better comprehend the required computing resources for training AI models and datasets. Liao et al. [34] proposed HPC ontology to capture HPC-related concepts using the web ontology language (OWL) [39]. HPC ontology allows users to query PLP and HPC information but is limited by the manual processes of data creation, collection, and updates. Moreover, HPC ontology depends on users to articulate their requirements, including conditions or constraints related to code, hardware, dataset, models, and more, resulting in a challenge to accurately capture and utilize such nuanced semantics.

In the second application, we evaluate data race detection for C/C++ and Fortran tasks using the public data race benchmark (DRB) [35]. Data race detection analyses can be broadly categorized into dynamic and static approaches. While dynamic approaches, like Intel Inspector [29], are commonly used, there is growing interest in static data race detection tools that complement dynamic approaches. LLOV [8] is a recent tool evaluated with DRB. However, designing such tools demands significant data race expertise and running code snippets, making the process time-consuming.

Leveraging LLMs for solving HPC tasks holds tremendous promise, as a single HPC-domain LLM can effectively handle various general applications, eliminating the need to devise separate methods for each task. However, training an LLM for the HPC domain [36, 37] presents non-trivial challenges, primarily revolving around the collection of domain-specific training data. In contrast to traditional language model training, which involves feeding large volumes of raw data like text, fine-tuning an LLM for HPC necessitates specialized instruction data (e.g., PromptInstructions [7] and Super-Natural-Instructions [49]). These instructions guide the model’s understanding of HPC concepts and tasks, ensuring it performs well in this domain. However, acquiring such instruction data proves to be a daunting task due to the associated costs and limitations in diversity. To address this issue, Wang et al. [46] propose a semi-automated process for instruction-tuning a pre-trained language model using instructional signals extracted from the model itself. This approach mitigates some of the challenges in data collection and provides an alternative means of training the LLM for HPC tasks. However, it requires a set of manually-written tasks to guide the overall generation process, which can be time-consuming and resource-intensive.

In this paper, we present HPC-GPT, a specialized LLM tailored for the HPC domain. To facilitate fine-tuning, we design an automatic collection method to gather instruction data for training the open-source base LLaMa [44] and LLaMa2 model [45]. Leveraging the power of the existing LLM (GPT-4), we generate domain-specific and contextually relevant instructions from unsupervised text. This approach ensures that the generated instructions are specific to the HPC domain and aligned with the provided unsupervised text. To address specific HPC tasks, we design distinct prompts that outline relevant requirements, enabling the model to generate instructions and answers separately. Subsequently, we establish filtering and pruning rules to eliminate instruction data that fails to meet the specified requirements. Our HPC-GPT model is built upon the foundation of the open-source LLaMa-13B base model, further enhancing its language generation capabilities. To showcase the potential of HPC-GPT, we utilize the two HPC applications mentioned earlier as illustrative examples to evaluate the performance. We summarize the main contributions of this paper as follows:

- Introduction of the HPC-GPT model, representing the first open-source LLM fine-tuned using HPC instruction data. This specialized model is specifically designed to excel in HPC tasks.
- Integration of HPC knowledge into the model, ensuring it possesses accurate and domain-specific information. By incorporating HPC knowledge, the model leads to enhanced performance in HPC applications.
- Release of two open-source datasets for the HPC domain: one containing HPC models and datasets, and the other dedicated to data race detection. These datasets provide valuable resources for researchers in the HPC community, facilitating further advancements in the field. The code 1 and datasets 2 are publicly available.
- Comparative evaluation against state-of-the-art methods. This progress underscores the model’s effectiveness and its potential to outperform existing approaches in HPC-related tasks.

2 RELATED WORK

2.1 Large Language Models

In the realm of language processing, recent strides in LLMs [3, 10, 17, 18] have showcased their superiority over earlier paradigms like pretraining [31] and fine-tuning [26]. This remarkable progress can be attributed to the substantial growth in model scale, resulting in qualitative shifts within LLMs, known as emergent abilities. These newfound capabilities encompass in-context learning, enabling the model to tackle zero-shot tasks, and the ability to follow chains of thought [48], thereby enhancing its performance on intricate tasks.

OpenAI’s groundbreaking work in developing ChatGPT [3] and GPT-4 [10] has brought about a paradigm shift in the understanding of LLMs. These models have demonstrated remarkable performance, yet OpenAI has maintained confidentiality regarding their training strategies and weight parameters. To address this limitation, LLaMa [44] and LLaMa2 [45] emerge as an open-source alternative to GPT, available in sizes ranging from 7 billion to 65 billion parameters. While LLaMa’s performance is comparable to GPT-3.5 in general tasks, it falls short in HPC-related tasks due to its training data being predominantly focused on general applications.

2. https://huggingface.co/datasets/HPC-GPT/HPC
2.2 Large Language Models for HPC

LLMs, such as GPT-4 and LLaMA, have been widely used in multiple domains, including natural language processing, visualization, and so on. However, applying them for analyzing and optimizing HPC tasks is still challenging due to the lack of HPC-specific support. To address this challenge, LM4HPC [12] represents the first attempt to adapt LLMs to the HPC domain. This is achieved by creating the LM4HPC framework to facilitate research into HPC analyses and optimizations using LMs. Tailored for supporting HPC datasets, AI models, and pipelines, the LM4HPC framework is built on top of a range of components from different levels of the machine learning software stack, with Hugging Face-compatible APIs. However, the framework currently is still relying on existing general LLM for HPC tasks which is suboptimal.

3 DESIGN OF HPC-GPT

In this section, we describe in detail the design of HPC-GPT.

3.1 HPC-GPT Overview

Figure 1 illustrates an overview of the proposed HPC-GPT framework, which can be divided into four main stages: HPC domain data collection, training, evaluation, and deployment. In the HPC domain data collection stage, we develop a data collection method that automatically gathers the necessary training data for two specific HPC applications. This data curation process ensures that the model is trained on relevant and domain-specific information. Moving on to the training stage, we employ supervised fine-tuning on the open-source LLM using the generated instruction data. Fine-tuning helps the model adapt to the intricacies of HPC-related tasks and enhances its performance in the targeted domain. In the evaluation stage, we rigorously assess the well-trained HPC-GPT model’s capabilities on a public data race detection dataset. This evaluation process validates the model’s effectiveness and performance in real-world scenarios. Finally, after successful training and evaluation, we deploy the HPC-GPT model to a web server, making it readily available for HPC scientists and researchers to utilize in their work.

3.2 Automatic Data Collection with LLM

```plaintext
/* The HPC knowledge is: */
/* (unsupervised knowledge data) */
/* According to the information above, please help me generate (number) questions. */
/* Here are the requirements: */
/* 1. Try not to repeat the verb for each question to maximize diversity. */
/* 2. Make sure the output is less than 50 words. */
/* 3. The questions can be asked under many conditions. */
/* 4. Do not generate the same or similar questions as generated before. */
/* Now, please generate the instructions following the above requirements. */
```

Listing 1: Instruction Generation Prompt

**Instruction Generation:** In this stage, we leverage the capabilities of the existing LLM, GPT-4, to generate domain-specific
instructions based on unsupervised text, encompassing both unstructured and structured HPC data. The unsupervised text serves as the input and includes a wide range of HPC-related knowledge data. To tailor the instructions for specific HPC tasks, we design instruction prompts with relevant requirements, as depicted in Listing 1. These prompts guide GPT-4 to respond with instructions that align with the provided text data. The generated instructions serve as guidance for the subsequent stages of the process.

The unsupervised knowledge data within the prompt consists of sequential text. Unstructured knowledge data, such as websites and conference papers, can be directly used after undergoing necessary cleaning processes. On the other hand, structured data, like tables, needs to be converted into unstructured textual data before it becomes usable. As illustrated in Figure 2, this conversion can be achieved through slot-filling using templates or by concatenating each data entry with its corresponding attribute name. These steps ensure that structured data is transformed into a format that can be effectively utilized by the language model for instruction generation and fine-tuning.

**Answer Generation:** In this stage, the language model is tasked with generating answers to the instruction questions based on the corresponding unsupervised HPC knowledge. The process can be expressed as $A = P(K, Q)$, where $K$, $Q$, and $A$ represent unsupervised knowledge, instruction question, and answer, respectively. Similar to the previous stage, the prompt for generating answers is provided, as shown in Listing 2. Using the instruction questions and the associated unsupervised knowledge, the language model is prompted to produce accurate and contextually relevant answers. This step enhances the model’s understanding of the provided knowledge and strengthens its ability to respond effectively to the instruction questions.

---

**Figure 2: Transformation of unsupervised structured data.**

**Listing 2: Instruction-Answer Generation Prompt**

Filtering and Pruning: Although we explicitly instruct the model with the prompt in Listing 2 to not generate the same or similar instructions and answers as generated before, we still observe that the model produces instruction data that violates these rules. Additionally, the generated instances of instructions also exhibit cases where they do not adhere to the required format and become unparseable. Therefore, it is necessary to further filter out these problematic examples. To mitigate these issues, we implement a postprocessing step to filter out inappropriate responses and correct any formatting errors. This involves developing heuristics and rule-based methods to identify and remove instances that violate the instructed constraints. By applying these filters, we ensure that the generated text adheres to the predefined guidelines and maintains the desired level of correctness.

**3.3 Training Pipeline**

There are two stages in the training pipeline: pretraining and supervised fine-tuning.

**3.4 Pretraining**

When training LLM for HPC domain, the base model provides foundational language understanding, enabling quicker adaptation to the new domain’s nuances. It offers transferable linguistic knowledge and reduces data requirements. The base model’s broad comprehension aids in fine-tuning while retaining the ability to generate accurate responses beyond the specific domain. We select LLaMA [44] and LLaMA 2 [44] as base models for pretraining. The reason is two-fold. Firstly, the LLaMA series are open-source and free to conduct research, in contrast to the commercial nature of the GPT series. Secondly, LLaMA outperforms other models, like GPT-3, in efficiency and resource utilization. LLaMA series is a collection of multi-lingual base models with parameters ranging from 7 billion to 70 billion. Here, We adopt the 13B version for both LLaMA and LLaMA 2 models, ensuring training accessibility and achieving superior performance.

**3.5 Supervised Fine-tuning**

Although LLMs exhibit remarkable performance in general domains, their lack of domain-specific knowledge results in suboptimal performance in HPC fields that require specialized expertise. The HPC field’s inherent nature necessitates models to possess comprehensive knowledge bases for relevant queries. Supervised fine-tuning has proven to be effective in tuning LLM for different tasks [40, 47]. Supervised fine-tuning helps the models perform satisfactorily under zero-shot scenarios with the cost of sufficient
annotated instructions. Inspired by the automatic construction of the instruction along with the instances (inputs and outputs) [46], we generate our instruction data based on the above HPC knowledge with the proposed method in Section 3.2. Table 1 shows the supervised fine-tuning data examples for 2 HPC tasks. The data instance has an instruction that describes the tasks in natural language. In HPC-based tasks, we consider the instructions and input are the same and leave the input null value.

4 EXPERIMENT

4.1 Experiments Setting

Our experimental setup takes place on a DGX A100 node within the ThetaGPU server, equipped with eight NVIDIA A100 Tensor Core GPUs and two AMD Rome CPUs, providing 320 GB GPU memory. For training HPC-GPT, we allocate 12 hours with 200 training epochs and set the learning rate to 2e-5. To optimize the process, we utilize a training and evaluation batch size of 16. To efficiently fine-tune HPC-GPT, we leverage fp16 precision to reduce memory requirements. Furthermore, we apply the LoRA [27] technique, known as low-rank adaptation, to reduce the number of trainable parameters. Additionally, we employ the PEFT [33] technique, which stands for parameter-efficient fine-tuning, to further enhance the performance of the pre-trained models.

4.2 HPC Unstructured and Structured Data

We collect the HPC raw data from GitHub, papers, and websites. For the PLP task, the unstructured data is from more than 40 papers related to PLP tasks, such as paper [23], and the structured data is from tables in the paper (e.g., [38]) and GitHub (e.g., [5]). For the MLPerf task, the unstructured data is from papers related to ML performance such as paper [42], and the structured data is from tables such as the results in the website [2]. We feed the raw data to LLM using the methods in Section 3.2 to collect the instruction data.

4.3 Two HPC Tasks

1. Managing AI Models and Datasets has two subtasks: 1) PLP task and 2) MLPerf task. PLP task refers to leveraging machine learning techniques to understand and analyze programming languages in a human-readable manner. A large number of different ML models and datasets are published each year related to different kinds of PLP tasks such as code generation [32], clone detection [32], source-to-source translation [30], defect correction [9], code documentation [28] and so on. It is difficult for people, especially newcomers, to identify representative ones to get started. Different model architectures are used to solve different types of PLP tasks, ranging from program understanding to code generation. It is a daunting job for people to pick the right architectures for a given task. The goal of this task is to facilitate the reuse of datasets and models related to PLP tasks, so researchers or developers can easily create customized ML pipelines to solve a given task. MLPerf [42] is a standardized benchmark designed to evaluate and compare the inference performance of machine learning models and frameworks. The benchmark includes different types of models and uses cases. Participants in the MLPerf benchmark must run their inference engines on a predefined set of models and datasets, following specified guidelines for hardware and software configurations. The goal of MLPerf task is to help the participants know the current exact system (e.g., 16-nodes-SPR-pytorch), processor (e.g., Intel(R) Xeon(R) Platinum 8462Y+), accelerator (e.g., NVIDIA H100-SXM5-80GB), and software (e.g., PyTorch NVIDIA Release 23.04) that can facilitate them to build similar ML models and datasets efficiently.

2. Data Race Detection is a critical task in HPC. It is aimed at finding data race bugs in multithreaded programs such as those using OpenMP. In general, a data race occurs when two or more threads perform conflicting accesses (with at least one access being a write) to a shared variable without any synchronization among the threads. In this task, we cover two programming languages (such as C/C++ and Fortran) using OpenMP. The goal of this task is to detect if there is a data race problem given the code snippet.

4.4 Baselines

HPC-GPT comprises two versions: HPC-GPT (L1) and HPC-GPT (L2), each integrating the respective base models, LLaMa and LLaMa 2. In order to demonstrate the superior performance of HPC-GPT, we conduct a comparative analysis with different baseline methods.

Task 1: Managing AI Models and Datasets

- ChatGPT is the state of the art of commercial large language model, we use GPT-4 as its model engine with superior performance.
- HPC Ontology [34] is a unified ontology for managing Al models and datasets for HPC.

Task 2: Data Race Detection

- LLOV [8] is a fast, lightweight, language agnostic, and static data race checker for OpenMP programs based on the LLVM compiler framework.
- Intel Inspector [29] is a dynamic analysis tool that detects threading and memory errors in C, C++ and Fortran codes. It supersedes Intel’s Thread Checker tool, with added memory error checking. Supported thread errors include race conditions and deadlocks.
- ROMP [23] is a tool for detecting data races in executions of scalable parallel applications that employ OpenMP for node-level parallelism.
- Thread Sanitizer [43] is a runtime data race detector developed by Google. ThreadSanitizer is now part of the LLVM and GCC compilers to enable data race detection for C++, Fortran and Go code.
- GPT-3.5 and GPT-4 are standard commercial LLM developed by OpenAI. GPT-4 with 1.7 trillion parameters is the latest and most advanced version of GPT compared with GPT-3 with 175 billion parameters.
- LLaMA [44] and LLaMA 2 [45] are standard open-source LLM developed by Meta. LLaMA 2 was trained on 40% more data than LLaMA. We use 13B versions for LLaMA and LLaMA 2 due to the tradeoff between performance and training costs.

4.5 Metrics

To evaluate HPC-GPT, we calculate counts of True Positive (TP), False Positive (FP), True Negative (TN), and False Negatives (FN) based on tool results and ground truth. We use five standard metrics:
Table 1: Instance with An Instruction

<table>
<thead>
<tr>
<th>Task1: Model and datasets for HPC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;instruction&quot;: &quot;What kind of dataset can be used if the language is C/C++ and the baseline is CodeBERT?&quot;,</td>
<td></td>
</tr>
<tr>
<td>&quot;input&quot;: &quot;&quot;,</td>
<td></td>
</tr>
<tr>
<td>&quot;output&quot;: &quot;The POJ-104 dataset can be used for clone detection tasks if the language is C/C++ and the baseline is CodeBERT.&quot;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 2: Data Race Detection</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;instruction&quot;: &quot;Given the code snippet: &quot;c\n#pragma omp parallel for\nfor (i = 1; i &lt;n; i++)\ny[i] = x[i] + y[i-1]\n\nhelp me detect if adding pragma will cause a data race problem? Answer 'yes' if it causes a data race problem and 'no' if it will not cause a data race problem.&quot;,</td>
<td></td>
</tr>
<tr>
<td>&quot;input&quot;: &quot;&quot;,</td>
<td></td>
</tr>
<tr>
<td>&quot;output&quot;: &quot;yes&quot;</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Dataset Information for Task 1

<table>
<thead>
<tr>
<th>Subtasks</th>
<th>Category</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>Performance Modeling</td>
<td>44</td>
<td>7.30%</td>
</tr>
<tr>
<td></td>
<td>Algorithm Classification</td>
<td>41</td>
<td>6.80%</td>
</tr>
<tr>
<td></td>
<td>Defect detection</td>
<td>47</td>
<td>7.79%</td>
</tr>
<tr>
<td></td>
<td>Clone detection</td>
<td>43</td>
<td>7.46%</td>
</tr>
<tr>
<td></td>
<td>Code Completion</td>
<td>39</td>
<td>6.47%</td>
</tr>
<tr>
<td></td>
<td>Compiler Analyses</td>
<td>37</td>
<td>6.14%</td>
</tr>
<tr>
<td></td>
<td>Code Repair</td>
<td>48</td>
<td>7.96%</td>
</tr>
<tr>
<td></td>
<td>Code Translation</td>
<td>41</td>
<td>6.80%</td>
</tr>
<tr>
<td></td>
<td>Close Testing</td>
<td>48</td>
<td>7.96%</td>
</tr>
<tr>
<td></td>
<td>Text-to-Code Generation</td>
<td>58</td>
<td>9.62%</td>
</tr>
<tr>
<td></td>
<td>Code Summarization</td>
<td>48</td>
<td>7.96%</td>
</tr>
<tr>
<td></td>
<td>Document Translation</td>
<td>52</td>
<td>8.62%</td>
</tr>
<tr>
<td>MLPerf</td>
<td>Code Search</td>
<td>53</td>
<td>9.12%</td>
</tr>
<tr>
<td></td>
<td>Submitter</td>
<td>324</td>
<td>17.80%</td>
</tr>
<tr>
<td></td>
<td>System</td>
<td>386</td>
<td>21.21%</td>
</tr>
<tr>
<td></td>
<td>Processor</td>
<td>347</td>
<td>19.07%</td>
</tr>
<tr>
<td></td>
<td>Accelerator</td>
<td>362</td>
<td>19.89%</td>
</tr>
<tr>
<td></td>
<td>Software</td>
<td>401</td>
<td>22.03%</td>
</tr>
</tbody>
</table>

Recall, Specificity, Precision, Accuracy, and F1 score to evaluate the quality of tools. We also report tool support rate (TSR), which is the ratio of how many test files are supported by a tool. The F1 score is a measure combining both precision and recall. It provides a single metric that weights precision and recall in a balanced way, requiring both to have higher values for the F1-score value to rise. The reported adjusted F1 score, F1 score multiplied by the TSR, can show the true ability of a tool.

4.6 Instruction Datasets for HPC

We have collected a total of 5.86k instruction data for two HPC applications. The dataset details for task 1 and task 2 are presented in Table 2 and Table 3, respectively. In Table 2, we focus on task 1: Managing AI models and datasets for PLP tasks and MLPerf tasks. There are 13 categories in PLP tasks covering various programming language processing subtasks, such as clone detection and code repair. The percentage range for these categories is 6.14% to 9.62%. Additionally, there are 5 categories dedicated to MLPerf tasks, such as the system and software, with a percentage range of 17.80% to 22.03%. In Table 3, the data race detection task has two main categories: "code snippet with data races" and "code snippet without data races." We have summarized 7 common data race types for both C/C++ and Fortran, including "missing data sharing" and "unresolvable dependencies." Additionally, we identified 7 types of data without data races, such as "Single thread execution" and "Use of synchronization." The percentage range for these categories is 5.96% to 8.17% for C/C++ and 6.15% to 8.25% for Fortran.

Overall, we have ensured that the number of different categories is balanced for various subtasks. This approach prevents one category from dominating and allows the LLM to maintain a well-rounded knowledge base for each task. This diverse and comprehensive dataset collection is crucial in fine-tuning the LLM effectively and preparing it to tackle a wide array of HPC-specific challenges with improved accuracy and performance.

4.7 Preliminary Results

4.7.1 Managing AI models and Datasets. This section presents some use cases using HPC-GPT for providing the dataset and model information and answering questions.

1) PLP Task. The AI models and dataset are fundamental information for PLP tasks [11, 13, 14]. We use natural language to describe the code translation-related questions in [23].

```
Question: "What kind of dataset can be used for code translation tasks if the source language is Java and the target language is C#?"
```

```
Answer (GPT-4): "For code translation tasks from Java to C#, you would need a dataset that consists of pairs of Java code and their corresponding equivalent C# code."
```

```
Answer (HPC-Ontology): "CodeTrans dataset"
```

```
Answer (HPC-GPT (L2)): "The CodeTrans dataset can be used for code translation tasks if the source language is Java and the target language is C#.
```

Listing 3: PLP Task Example

2) MLPerf Task. The system, processor, and software-related details are important for MLPerf tasks. We use natural language to describe the MLPerf-related questions in [2].

```
Question: "What is the System if the Accelerator used is NVIDIA H100-SXM5-80GB and the Software used is MXNet?
```

```
NVIDIA H100-SXM5-80GB and the Software used is MXNet NVIDA Release 23.04."
```
Table 3: Dataset Information for Task 2

<table>
<thead>
<tr>
<th>Subtasks</th>
<th>Metric</th>
<th>Code snippet with data races</th>
<th>Code snippet without data races</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Unresolvable dependencies</td>
<td>Missing data sharing clauses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Missing synchronization</td>
<td>SIMD data races</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accelerator data races</td>
<td>Undefined behavior</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Numerical kernel data races</td>
<td>Single thread execution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Use of data sharing clauses</td>
<td>Use of synchronization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Use of SIMD directives</td>
<td>Use of accelerator directives</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Use of special language features</td>
<td>Numerical kernels</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subtasks</th>
<th>Tools</th>
<th>Version</th>
<th>Compiler</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/C++</td>
<td>Number</td>
<td>122</td>
<td>GCC/gfortran 7.4.0</td>
<td>7.55%</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
<td>122</td>
<td>GCC/gfortran 7.4.0</td>
<td>7.55%</td>
</tr>
<tr>
<td></td>
<td>Fortran</td>
<td>Number</td>
<td>123</td>
<td>Fortran 2021.3.0</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
<td>123</td>
<td>Fortran 2021.3.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.93%</td>
<td>6.34%</td>
<td>7.42%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.74%</td>
<td>8.41%</td>
<td>6.91%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.04%</td>
<td>6.21%</td>
<td>8.08%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.66%</td>
<td>8.25%</td>
<td>6.15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.85%</td>
<td>7.86%</td>
<td>6.35%</td>
</tr>
</tbody>
</table>

Table 4: Data Race Detection Tool and Compiler Version

<table>
<thead>
<tr>
<th>Tools</th>
<th>Version</th>
<th>Compiler</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThreadSanitizer</td>
<td>10.0.0</td>
<td>Clang/LLVM 10.0.0</td>
</tr>
<tr>
<td>Intel Inspector</td>
<td>2021.1</td>
<td>Intel Compiler 2021.3.0</td>
</tr>
<tr>
<td>ROMP</td>
<td>2024.0</td>
<td>GCC/gfortran 7.4.0</td>
</tr>
<tr>
<td>LLOV</td>
<td>N/A</td>
<td>Clang/LLVM 6.0.1</td>
</tr>
</tbody>
</table>

Listing 4: MLPerf Task Example

Results. The results for the PLP task and MLPerf task are shown in Listing 3 and 4 respectively. Upon analyzing the outputs, it is evident that HPC-GPT demonstrates its superiority in handling these tasks compared to GPT-4. In the PLP task, when the user asks, "What kind of dataset can be used for code translation tasks if the source language is Java and the target language is C#?", HPC-GPT is capable of providing an exact answer with the "CodeTrans dataset", while GPT-4 merely repeats the question due to its lack of relevant knowledge. Similarly, in the MLPerf task, when the user presents a system-related question such as, "What is the system if the Accelerator used is NVIDIA H100 - SXM5 - 80 GB and the Software used is MXNet NVIDIA Release 23.04?", HPC-GPT promptly offers a specific answer, "dgxh100_n64", demonstrating its understanding of the question. Conversely, ChatGPT's response includes a general introduction of the Accelerator but fails to provide the correct answer, indicating its lack of relevant knowledge regarding the MLPerf task.

To address this limitation, HPC-Ontology can also deliver accurate answers by leveraging SPARQL query language. However, it requires manual effort to write SPARQL queries for different questions and answers, making it less scalable. HPC-GPT overcomes this challenge by effectively translating human language into embeddings, enabling it to process all the information and provide relevant answers to various queries. In both tasks, HPC-GPT demonstrates its versatility in handling different conditions and benefits users by allowing them to express their questions freely while receiving accurate and related answers.

4.7.2 Data Race Detection. In this section, we present the evaluation results of HPC-GPT on DataRaceBench V1.4.0 [15, 35] with 177 C/C++ test programs and 166 Fortran test programs. Among these, 88 C/C++ and 84 Fortran test cases exhibit data races, while 89 C/C++ and 82 Fortran test cases are free from data races. We utilize the metrics defined in Section 4.5 to assess the performance of data race detection tools (LLOV, Intel Inspector, ROMP, and Thread Sanitizer) and LLM-based methods (GPT-3.5, GPT-4, LLaMa, LLaMa2, HPC-GPT (L1) and HPC-GPT (L2)). The compiler version is shown in Table 4. The results are summarized in Table 5, where the best result for each metric is highlighted in bold. In C/C++ language, Thread Sanitizer excels in Adjusted F1 (0.8679), specificity (0.9888), precision (0.9857), and accuracy (0.8826). Remarkably, HPC-GPT (L2) secures second place in accuracy (0.8037), Adjusted F1 (0.8072), precision (0.9857), and accuracy (0.8826). HPC-GPT (L2) leads in Recall (0.8433) and Adjusted F1 (0.8333). Compared with LLM-based methods, both HPC-GPT (L1) and HPC-GPT (L2) demonstrate notable improvements over LLaMa, LLaMa 2, GPT-3.5, GPT-4 across five key metrics (recall, specificity, precision, accuracy, and adjusted F1). In C/C++ language, HPC-GPT (L2) achieves improvements of 36.11%, 34.84%, 26.33%, 11.1%, and 3.85% compared with LLaMa, LLaMa 2, GPT-3.5, GPT-4, and HPC-GPT (L1). For Fortran, HPC-GPT (L2) attains enhancements of 31.89%, 35.23%, 21.34%, 15.79%, and 7.28% over LLaMa, LLaMa 2, GPT-3.5, GPT-4, and HPC-GPT (L1). Notably, all LLM-based methods share the same TSR due to an 8k token constraint, limiting input length. For C/C++, TSR is lower than existing tools, with 14 test cases exceeding 8k tokens. Conversely, Fortran’s TSR for LLM-based methods is 1.0, surpassing existing tools.

Answer (GPT-4): "As of my latest update in September 2021, the NVIDIA H100-SXM5-80GB is a data center GPU designed for high-performance computing and deep learning workloads."

Answer (HPC-Ontology): "dgxh100_n64"

Answer (HPC-GPT (L2)): "If the Accelerator used is NVIDIA H100-SXM5-80GB and the Software used is MXNet NVIDIA Release 23.04, the System is dgxh100_n64."
### Table 5: Results of Individual Data Race Detection Tools and LLM-Based methods.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Language</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>Recall</th>
<th>Specificity</th>
<th>Precision</th>
<th>Accuracy</th>
<th>TSR</th>
<th>Adjusted F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLOV</td>
<td>C/C++</td>
<td>55</td>
<td>9</td>
<td>78</td>
<td>29</td>
<td>0.6666</td>
<td>0.8965</td>
<td>0.8656</td>
<td>0.7816</td>
<td>0.9613</td>
<td>0.7532</td>
</tr>
<tr>
<td>Intel Inspector</td>
<td></td>
<td>76</td>
<td>41</td>
<td>45</td>
<td>10</td>
<td>0.8373</td>
<td>0.5287</td>
<td>0.6495</td>
<td>0.7052</td>
<td>0.9558</td>
<td>0.7487</td>
</tr>
<tr>
<td>ROMP</td>
<td></td>
<td>63</td>
<td>12</td>
<td>65</td>
<td>18</td>
<td>0.775</td>
<td>0.8333</td>
<td>0.8266</td>
<td>0.8037</td>
<td>0.8729</td>
<td>0.8000</td>
</tr>
<tr>
<td>Thread Sanitizer</td>
<td></td>
<td>69</td>
<td>1</td>
<td>89</td>
<td>20</td>
<td>0.7752</td>
<td>0.9888</td>
<td>0.9857</td>
<td>0.8826</td>
<td>0.9899</td>
<td>0.8679</td>
</tr>
<tr>
<td>GPT-3.5</td>
<td></td>
<td>52</td>
<td>7</td>
<td>45</td>
<td>30</td>
<td>0.6341</td>
<td>0.5555</td>
<td>0.5909</td>
<td>0.5951</td>
<td>0.9209</td>
<td>0.6117</td>
</tr>
<tr>
<td>GPT-4</td>
<td></td>
<td>65</td>
<td>31</td>
<td>50</td>
<td>17</td>
<td>0.7926</td>
<td>0.6172</td>
<td>0.6770</td>
<td>0.7055</td>
<td>0.9209</td>
<td>0.7303</td>
</tr>
<tr>
<td>LLaMa</td>
<td></td>
<td>65</td>
<td>21</td>
<td>20</td>
<td>17</td>
<td>0.7926</td>
<td>0.2469</td>
<td>0.5158</td>
<td>0.5247</td>
<td>0.9209</td>
<td>0.625</td>
</tr>
<tr>
<td>LLaMa2</td>
<td></td>
<td>71</td>
<td>66</td>
<td>15</td>
<td>11</td>
<td>0.8658</td>
<td>0.1851</td>
<td>0.5184</td>
<td>0.5276</td>
<td>0.9209</td>
<td>0.6484</td>
</tr>
<tr>
<td>HPC-GPT (L1)</td>
<td></td>
<td>64</td>
<td>20</td>
<td>61</td>
<td>18</td>
<td>0.7804</td>
<td>0.7530</td>
<td>0.7619</td>
<td>0.7668</td>
<td>0.9209</td>
<td>0.7710</td>
</tr>
<tr>
<td>HPC-GPT (L2)</td>
<td></td>
<td>67</td>
<td>17</td>
<td>64</td>
<td>15</td>
<td>0.8171</td>
<td>0.7901</td>
<td>0.7976</td>
<td>0.8037</td>
<td>0.9209</td>
<td>0.8072</td>
</tr>
</tbody>
</table>

#### 5 CHALLENGES AND POSSIBLE SOLUTIONS

**The Token Length of Existing LLMs:** The limitation of LLM-based methods in detecting data races that exceed the 8k token limit poses a significant challenge in practical applications. When dealing with real-world software projects, data races may involve extensive sections of code that surpass the token limit. As a consequence, LLM-based methods fail to process such data race instances, leading to incomplete and potentially inaccurate results. To address this limitation, one approach is to investigate ways to extend the token limit of LLM models, allowing them to handle longer code snippets effectively. This involves optimizing the model architecture or leveraging advanced hardware configurations. Another avenue for improvement involves devising a pre-processing or partitioning mechanism to break down large code snippets into smaller, manageable segments that fit within the token limit. This way, LLM-based methods can analyze each segment individually and then combine the results to detect data races across the entire codebase.

**How to update HPC-GPT with Latest Data:** Updating HPC-GPT with the latest data poses a challenge due to the continuous release of datasets and models. Several strategies can be employed for effective updates. Initially, new data can be periodically gathered to retrain the entire HPC-GPT alongside existing data. Another method involves creating a checkpoint of the current model version and then resuming training using the newly acquired data. Another approach leverages the LangChain framework [1], wherein HPC-GPT integrates new data seamlessly. The LongChain APIs enable the storage of text within semantic vector stores. This integration process entails the division of text into chunks, followed by embedding and matching prompts with the most relevant vector chunks. Consequently, this enhances the context of responses while adhering to token limitations.

#### 6 CONCLUSION

This paper proposes HPC-GPT, a large language model specifically designed for the HPC domain. We explore the capabilities of this model in addressing two common tasks in HPC and demonstrate its excellent performance. In the first task, concerning model and dataset selection for HPC, HPC-GPT exhibits the ability to retrieve and extract HPC-related datasets and models based on human expressions. It efficiently handles multiple conditions, providing valuable support for researchers in the HPC field. Additionally, we evaluate HPC-GPT’s performance in data-race detection in an OpenMP program, a critical concern in parallel computing. Our experiments demonstrate that HPC-GPT achieves good performance compared to existing data-race detectors.

### ACKNOWLEDGMENTS

This research was funded in part by and used resources at the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. This work is also prepared by LLNL under Contract DE-AC52-07NA27344 (LLNL-CONF-853156) and supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research.

### REFERENCES

[2] 2023. MLPerf Training v3.0 Results. https://docs.google.com/spreadsheets/d/1sFbua0hEqvQwOla5XHk89JVR0ZtH0T_CaAL3mK/pubhtml?gid=0&single=false&widget=false&headers=false&chrome=true


