

# From Legacy Fortran to Portable Kokkos: An Autonomous Agentic AI Workflow

Sparsh Gupta, Kamalavasan Kamalakkannan, Maxim Moraru, Galen Shipman, and Patrick Diehl

**Abstract**—Scientific applications continue to rely heavily on legacy Fortran codebases originally developed for homogeneous, CPU-based systems. As High-Performance Computing (HPC) evolves towards heterogeneous GPU-accelerated architectures, many modern accelerators lack native Fortran bindings, creating an urgent need to translate and optimize legacy code for portability. Frameworks like Kokkos provide performance portability across multiple architectures and a single-source C++ abstraction, but manual porting from Fortran to Kokkos requires significant domain expertise and remains time-intensive. While large language models (LLMs) have demonstrated promise in source-to-source code generation, their use in building fully autonomous workflows for translating and optimizing parallel code is largely unexplored, particularly in the context of achieving performance portability across diverse hardware. This paper presents an agentic AI workflow in which specialized LLM “agents” collaborate to translate, validate, compile, run, test, debug, and optimize Fortran kernels into portable Kokkos C++ programs. Our results show that the pipeline successfully modernizes a variety of benchmark kernels, producing performance-portable Kokkos codes across hardware partitions. Paid OpenAI models such as GPT-5 and o4-mini-high executed the full workflow for only a few U.S. dollars, producing optimized codes that exceeded the Fortran baselines, whereas open-source models like Llama4-Maverick often failed to produce functional codes. This work demonstrates the feasibility of agentic AI for Fortran-to-Kokkos code transformation and offers a path toward autonomously modernizing legacy scientific applications to run portably and efficiently across a diverse set of supercomputers. It further illustrates the potential of LLM-driven agentic systems to perform structured, domain-specific reasoning tasks in scientific and systems-oriented applications.

**Impact Statement**—Many of today’s most important scientific applications still rely on decades-old Fortran code, which was never designed for today’s GPU-driven supercomputers. This creates a major barrier to accelerating nuclear simulations, astrophysical modeling, materials science, drug discovery, etc. that depend heavily on high-performance computing. Our work demonstrates that artificial intelligence agents can autonomously translate and optimize these legacy codes into modern, portable

C++ programs that run efficiently on diverse architectures. What once required weeks of expert programmer time and significant costs can now be achieved in just a few hours with paid large language models, at a cost of only a few U.S. dollars and without human intervention. By lowering the expertise, time, and cost barriers to large-scale code modernization, this approach can accelerate scientific discovery, extend the lifetime of critical legacy software, and expand access to next-generation supercomputing, ensuring both science and industry can adapt rapidly to future hardware capabilities.

**Index Terms**—Agentic AI, Fortran, Generative AI, High-Performance Computing (HPC), Large Language Models (LLMs)

## I. INTRODUCTION

LEGACY Fortran code remains the backbone of many crucial scientific computing applications across national laboratories and academic institutions, such as the WRF model [1], Code\_saturne [2], CHARMM [3], *etc.* These Fortran codes were originally designed for homogeneous, CPU-based systems and have been performance-tuned over decades for intra-node parallelism using interfaces like MPI [4] and OpenMP [5]. However, with the transition to heterogeneous High-Performance Computing (HPC) systems featuring GPU-accelerated hardware from vendors like NVIDIA, AMD, and Intel, there is a growing need to port legacy codes to run efficiently on modern architectures. Many of these accelerators lack native Fortran bindings [6], creating significant barriers to portability and performance. To address this, frameworks such as the Kokkos C++ Performance Portability Ecosystem [7] have emerged. Kokkos was originally developed under the U.S. Department of Energy’s Exascale Computing Project and is now part of the Linux Foundation’s High Performance Software Foundation. It has become a widely adopted tool for writing portable HPC applications as it offers performance portability by allowing developers to write single-source C++ code that can run efficiently with multiple programming models like CUDA [8], HIP [9], SYCL [10], OpenMP [5], HPX [11], and C++ threads. The transition to Kokkos seems very promising, however, the practical process of manually porting Fortran code to Kokkos requires a significant expertise in C++ with a deep understanding of Kokkos for parallel programming on HPC architectures. This process can be exceptionally time-consuming and requires tedious fine-tuning and debugging, creating a major bottleneck in modernizing scientific code.

Large Language Models (LLMs) have recently emerged as powerful tools for code translation, code generation [12], and even coding agents like OpenAI’s Codex [13]. Models

Research presented in this article was supported by the National Security Education Center (NSEC) Informational Science and Technology Institute (ISTI) using the Laboratory Directed Research and Development program of Los Alamos National Laboratory project number 20240479CR-IST. This research used resources of the National Energy Research Scientific Computing Center, a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. This work was also supported by the U.S. Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001). LA-UR-25-28882

Sparsh Gupta is with the Los Alamos National Laboratory, Los Alamos, NM 87544 USA and the Franklin W. Olin College of Engineering, Needham, MA 02492 USA (e-mail: sgupta1@olin.edu).

Kamalavasan Kamalakkannan, Maxim Moraru, Galen Shipman, and Patrick Diehl are with the Los Alamos National Laboratory, Los Alamos, NM 87544 USA. (e-mails: {kamalavasan,moraru,gshipman,diehlpk}@lanl.gov)

such as CodeLlama [14] and GPT-4 [15] have shown strong performance on coding tasks ranging from code infilling to code repair [16]. These recent developments demonstrate the capacity of LLMs to handle code-related tasks effectively. However, existing methods have not fully explored their potential in integrating diagnostics, iterative performance tuning, and hardware-specific optimizations, particularly in HPC contexts where performance portability is essential. Achieving this integration motivates our exploration into structured, multi-agent AI workflows for modernizing legacy Fortran codebases.

Therefore, in this work, we introduce a fully automated agentic AI workflow for translating and optimizing Fortran kernels to Kokkos C++ across different architectures. This pipeline consists of specialized LLM agents that collaborate to translate source code, validate it, apply fixes, manage compilation and execution on HPC clusters, identify and fix build/runtime failures, test functionality and debug if needed, and finally, propose optimizations based on metrics and outputs from hardware GPU profilers. These agents leverage LLMs obtained through the OpenAI API as well as open-source LLMs. This workflow is evaluated on benchmark Fortran kernels from the NAS Parallel Benchmarks [17] and OpenBLAS (DGEMM) [18]. Our evaluation shows that the workflow produces functionally correct and performance-portable Kokkos implementations of benchmark kernels, with OpenAI models such as GPT-5 [19] and o4-mini-high [20] being able to execute the full pipeline across all kernels and hardware partitions. In contrast, the open-source Llama4-Maverick [21] model often failed to complete the workflow, highlighting the current gap between proprietary and open-source LLMs in this domain. One likely explanation is the difference in model size. Llama4-Maverick was trained as a 400B parameter mixture-of-expert model with 17B active parameters per token. On the other hand, OpenAI doesn't disclose this information but we assume their LLMs have significantly larger number of parameters.

The rest of the paper is structured as follows. Section II reviews related work on Fortran modernization and LLM-driven code generation. Section III provides background on Kokkos, SLURM, Spack, and Agentic AI. Section IV details the benchmark Fortran kernels used for evaluation. Section V outlines our proposed methodology and workflow. The experimental setup is presented in Section VI, followed by the results in Section VII. Finally, Section VIII concludes the paper and discusses future work.

## II. RELATED WORK

Researchers have been exploring translation of Fortran codebases for decades. Early automated tools like *f2c* [22] enabled basic translation from Fortran to C, facilitating reuse of legacy code in more contemporary environments. More sophisticated source-to-source compilers such as LLNL's ROSE framework [23] have been used to refactor Fortran HPC applications and enabling semi-automatic parallelization by injecting parallel constructs (e.g., OpenMP or OpenACC pragmas) into legacy Fortran loops. *LFortran* is a newer open-source Fortran compiler built on LLVM [24], but it remains in alpha

development and is not yet capable of translating most third-party scientific codes. While it supports interactive execution and offers features such as abstract syntax tree (AST) manipulation and experimental refactoring, many Fortran language features and compiler capabilities remain incomplete. As such, *LFortran* is currently not a practical tool for modernizing large-scale legacy applications, and performance optimization is far from being within its scope.

In parallel, large language models (LLMs) have started to evolve from single-step code generation towards structured, agentic workflows that iteratively improve, debug, and optimize software code. One such approach is exemplified in *ChatDev* [25], where it leverages multiple LLM agents that interact through natural-language dialogue to collaboratively complete full-cycle software development tasks. Similarly, *MetaGPT* [26] integrates human-like procedures into specialized agent prompts, enabling multiple role-based LLM agents to sequentially refine code implementations while acting like a simulated software company, improving task performance over unstructured prompting. Further, *CodeChain* [27] has demonstrated that using a chain-of-thought methodology to incrementally generate and revise modularized code achieves notable accuracy improvements on competitive programming benchmarks. Additionally, [28] demonstrated a pipeline in which LLM agents iteratively generate, execute, and debug code using runtime feedback. Collectively, these works underscore the potential of structured, iterative, and multi-agent workflows to substantially enhance LLM-driven code generation pipelines.

Beyond agentic systems, recent benchmarking has also assessed the standalone coding abilities of LLMs. For example, Llama2-70B could generate correct code and tests for simpler tasks but struggled with complex parallel workloads, often requiring manual intervention [29]. Additional works have explored earlier-generation LLMs (e.g., Codex, GPT-3, Llama-2) for generating HPC kernels and BLAS routines using manual or one-shot prompting [30], [31], [32]. Similarly, the ParEval benchmark systematically evaluated LLMs on parallel code generation tasks spanning models such as OpenMP, MPI, CUDA, HIP, and Kokkos, across 420 scientific problems [33]. Such studies highlight both the promise and the current limitations of applying LLMs directly to code generation without structured workflows. Building upon this, recently, a study on LLM-assisted translation of Fortran to C++ using open-weight models on multiple hardware platforms was conducted in [34]. The study quantified compilation success rates, output similarity to human-translated code, and functional correctness, demonstrating promising translation success but highlighting that many translations failed to compile or required manual post-translation debugging. Similarly, *Fortran2C++* [35] provided a multi-turn dialogue dataset and pipeline that leverages a dual-agent dialogue system to iteratively refine Fortran-to-C++ translations, improving compile success and CodeBLEU scores. Most recently, [36] explored the use of LLMs for translating Fortran HPC kernels (OpenMP/OpenACC) to C/C++ targeting multiple backends including OpenMP, OpenACC, CUDA, HIP, and Kokkos. Their method relies on multimodal prompting and fine-tuning to improve translation accuracy,

and requires manual correction and tuning during the process. These LLM-based workflows indicate strong potential for Fortran code translation, but most efforts focus primarily on syntactic and semantic correctness and rely on single-turn or manually guided prompting. As a result, they do not address autonomous multi-stage compilation, execution, debugging, or iterative performance optimization across architectures. In contrast, our work focuses specifically on Fortran-to-Kokkos modernization and evaluates a fully autonomous, multi-agent workflow without any human intervention or model fine-tuning. Our goal is to assess whether modern LLMs, when orchestrated through agentic workflows, can autonomously deliver functionally correct and GPU-optimized Kokkos code, emphasizing automated optimization and scalability of the modernization pipeline rather than one-shot translation accuracy.

Despite significant progress in both Fortran code modernization tools and recent LLM-based agentic workflows, important gaps persist in this field. Traditional source-to-source translation tools, ROSE and LFortran, primarily handle modernization at compile time but do not engage with systematic performance optimization. Similarly, emerging LLM-driven workflows have predominantly focused on correctness and syntactic accuracy, neglecting crucial aspects such as performance, hardware-specific optimizations and portability, and automated refinement based on feedback. This paper specifically addresses these limitations by introducing a novel, end-to-end agentic AI workflow. Our approach integrates translation with systematic compilation, execution monitoring, performance profiling, and iterative optimization stages, facilitating autonomous modernization of legacy Fortran kernels into high-performance, architecture-portable Kokkos C++ programs.

### III. BACKGROUND

#### A. Kokkos

Kokkos [7] allows developers to write single-source C++ code, and its design abstracts both the execution space (*e.g.*, GPU threads vs. CPU cores) and the memory space (*e.g.*, device vs. host memory). This enables portable parallel code that can be compiled and executed across diverse target architectures without requiring changes to the underlying logic. At its core, Kokkos provides high-level abstractions for parallel patterns, including `parallel_for` for launching parallel loops, `parallel_reduce` for reduction operations, and `parallel_scan` for prefix computations. It also supports *hierarchical parallelism*, such as thread teams (which allow a collection of threads to synchronize and share a “scratch pad” memory) and vector lanes (where multiple data elements are processed simultaneously by a single instruction). Additionally, Kokkos includes policies for specifying how the data structures are laid into the memory with patterns for common data structures. Beyond the core programming model, Kokkos includes components such as Kokkos Kernels, a library of portable implementations for linear algebra and graph algorithms widely used in HPC applications, and Kokkos Tools, which provides integration with performance profilers and

debugging tools. These features distinguish Kokkos from other programming models by offering a more complete ecosystem for achieving performance portability and leveraging GPU architectures effectively.

#### B. SLURM

SLURM [37] (Simple Linux Utility for Resource Management) is an open-source workload manager used for orchestrating job scheduling on high-performance computing (HPC) clusters. It is widely regarded as the state-of-the-art resource manager on supercomputers and is used to allocate computational resources and manage job queues, supporting batch and interactive job submission, array jobs, resource constraints, and dependency chains, making it well-suited for our workflow. Although our pipeline is currently implemented on SLURM, its modular design allows easy adaptation to other schedulers if needed.

#### C. Spack

Spack [38] is an open-source package manager widely adopted in HPC for managing software stacks. Spack is already deployed at most supercomputing centers and integrates with job schedulers and module systems, enabling reproducible, portable environments across architectures. It allows multiple versions and build configurations of the same package to coexist, which is critical for scientific codes with diverse dependencies. In our workflow, it is used to manage compilers, Kokkos builds on different backends, etc.

#### D. Agentic AI

Agentic AI refers to an artificial intelligence (AI) system composed of multiple large language model (LLM) agents that operate autonomously, with minimal intervention, to accomplish complex tasks through structured-decision making, tool use, and inter-agent communication. Unlike conventional one-shot LLM prompting, agentic systems maintain state, iterate over failures, support structured outputs, and decompose problems into subgoals delegated to specialized agents. An “agent”, in the context of agentic AI, is an autonomous software component powered by an LLM and equipped with a set of instructions, tools, memory, and contextual decision-making capabilities to perform a specific task or role. Each agent typically receives structured input, maintains contextual state, calls external tools or functions as needed, and produces structured output. Agents can be general-purpose or specialized, and are often designed to collaborate with other agents by passing intermediate outputs or task assignments. This allows for complex workflows to be decomposed into more easily solvable, breakable subtasks.

Several frameworks such as LangChain, CrewAI, Microsoft AutoGen [39], etc. have recently emerged to build agentic AI systems, providing abstractions for tool integration, planning, and agent coordination. In this work, we use the open-source framework OpenAI Agents SDK [40], which provides a Python-based lightweight, extensible framework for building multi-agent LLM workflows. It supports wrapping Python

functions through function tools, enabling agents to execute tasks programmatically with external systems such as compilers, file systems, job schedulers, and profilers. Each agent can be configured with structured prompts, internal memory, and access to task-specific tools. A key advantage of the SDK is its flexibility in model integration so that developers can easily route requests through different LLM providers, including OpenAI models served via APIs, or open-source models through self-hosted endpoints. It also supports inter-agent communication and task-delegation, allowing agents to operate in a coordinated manner by sharing intermediate outputs or feedback. These capabilities make the OpenAI Agents SDK particularly useful for orchestrating complex, tool-driven AI workflows.

#### IV. BENCHMARK FORTRAN KERNELS FOR EVALUATION

In this work, we evaluate our pipeline on the following Fortran 90 kernels, primarily chosen for their parallel complexity and diverse functionalities. To ensure consistent benchmarking with the generated Kokkos code, we modularized each benchmark kernel as a standalone Fortran subroutine with pre-initialized values that accepts two input parameters (same as the Kokkos programs): the problem size ( $n$ ) and the number of kernel repetitions. These modularized subroutines were then wrapped in minimal driver programs for execution, testing, and runtime benchmarking separately. Note that we only pass the modularized subroutines (without the driver program) to the agentic AI workflow for translation. A summary of the selected kernels is provided in Table I.

TABLE I  
BENCHMARKS USED IN THIS WORK

Kernel	Type	Modularized Subroutine Lines of Code
CG	Memory-bound	165
EP	Compute-bound	129
MG	Memory-bound	139
FT	Memory-bound	230
DGEMM	Compute-bound	177

##### A. STREAM Kernels

We tested our pipeline on the STREAM benchmark kernels [41] [42] (*Copy*, *Scale*, *Add*, and *Triad*) before evaluating the larger kernels. These kernels are simple and we did not use them for any quantitative evaluation or results, but only for testing while development of the workflow to ensure the pipeline functioned correctly on small kernels.

##### B. NAS Parallel Benchmarks (NPB)

The NPB3.4 [17] are a suite of performance kernels developed by the NASA Advanced Supercomputing (NAS) Division to evaluate the performance of highly parallel supercomputers.

1) *Conjugate Gradient (CG)*: The CG benchmark estimates the smallest eigenvalue of a large sparse symmetric positive-definite matrix. It employs an inverse iteration algorithm, utilizing the conjugate gradient method iteratively to solve linear systems of the form  $Ax = b$ , where  $A$  is a sparse symmetric positive-definite matrix,  $x$  is the unknown vector, and  $b$  is a known vector. CG is representative of unstructured memory access, indirect indexing, and irregular loop structures.

2) *Embarrassingly Parallel (EP)*: The EP kernel generates independent Gaussian random variates using the Marsaglia polar method and estimates values such as the integral of the Gaussian probability distribution.

$$x_1 = \sqrt{\frac{-2 \ln(u_1)}{u_1^2 + u_2^2}} u_1, \quad x_2 = \sqrt{\frac{-2 \ln(u_1)}{u_1^2 + u_2^2}} u_2, \quad (1)$$

where  $u_1, u_2$  are uniformly distributed random numbers within the range  $(-1, 1)$ , and  $x_1, x_2$  are resulting Gaussian-distributed random variables. This kernel involves no inter-task communication and is representative of purely parallel, compute-intensive workloads with minimal memory dependencies.

3) *Multi-Grid (MG)*: The MG kernel approximates the solution to a three-dimensional discrete Poisson equation using a V-cycle multigrid method, represented by the following PDE:

$$\nabla^2 u(x, y, z) = v(x, y, z), \quad (2)$$

where  $\nabla^2$  denotes the 3D Laplacian operator, and  $u$  and  $v$  represent the unknown solution and given source terms, respectively. The multigrid method iteratively approximates this solution across multiple resolution levels (grids). This kernel involves regular memory access patterns and hierarchical computation, making it representative of PDE solvers and multi-resolution grid algorithms. The kernel is memory-intensive and involves both short- and long-distance communication across multiple mesh levels.

4) *Fourier Transform (FT)*: The FT kernel solves a three-dimensional partial differential equation (PDE) using the Fast Fourier Transform (FFT), computed as:

$$F(k_x, k_y, k_z) = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} \sum_{z=0}^{N_z-1} f(x, y, z) e^{-2\pi i \left( \frac{k_x x}{N_x} + \frac{k_y y}{N_y} + \frac{k_z z}{N_z} \right)}, \quad (3)$$

where  $F(k_x, k_y, k_z)$  represents the transformed frequency-domain coefficients of the spatial domain data  $f(x, y, z)$ , and  $N_x, N_y, N_z$  denote grid sizes along each dimension. This kernel consists of structured memory access and global communication, characteristic of many PDE solvers and signal-processing workloads.

##### C. DGEMM from OpenBLAS

DGEMM (Double-precision GEneral Matrix Multiplication) is a fundamental routine selected from the OpenBLAS [18] library, specifically designed for performing matrix-matrix multiplication with double-precision floating-point numbers:

$$C = \alpha AB + \beta C, \quad (4)$$

where  $A, B, C$  are matrices, and  $\alpha, \beta$  are scalar multipliers. This kernel is ideal for evaluation as it is both compute-bound and cache-intensive, with performance strongly tied to memory layout, loop ordering, and tiling.



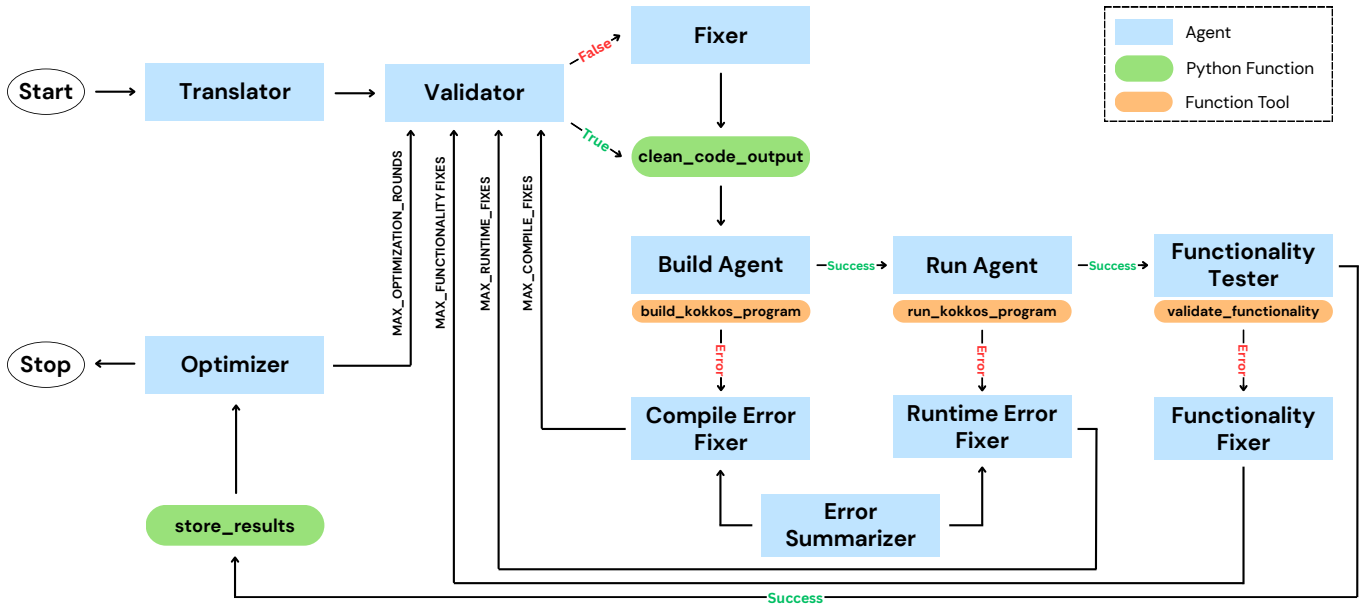


Fig. 1. Agentic AI workflow for autonomous Fortran-to-Kokkos translation, validation, compilation, runtime execution, functionality testing, and performance optimization. Fixer Agents are triggered on error events (e.g., failed compilation, runtime fault, or incorrect functionality testing output). Agent invocation limits at each stage are enforced via configurable thresholds (e.g., MAX\_COMPILE\_FIXES). Function tools invoked by the Build and Run agents (build\_kokkos\_program, run\_kokkos\_program) utilize SLURM to schedule and monitor jobs and Spack to load the correct Kokkos environment on the hardware partitions. All artifacts and metrics are versioned and stored per version run using the store\_results function.

## V. METHODS

### A. Overview of the Agentic AI Workflow

The proposed workflow is a fully autonomous pipeline for translating, building, running, testing, and optimizing legacy HPC Fortran kernels into performance-portable Kokkos C++ programs. The workflow is composed of a set of specialized LLM agents and is designed to operate without manual intervention and includes mechanisms for handling build failures, runtime errors, and functional correctness mismatches through iterative agent invocations. A high-level schematic of the complete workflow is shown in Fig. 1.

### B. Agents and Roles

The core of the proposed workflow is composed of modular, role-specific agents that operate sequentially within each stage of the pipeline, collaborating via intermediate outputs to progressively refine and optimize the generated code. We group the agents into four phases of the pipeline, as outlined below.

1) *Translation and Validation*: The pipeline begins with a *Translator Agent*, responsible for converting legacy Fortran kernels into standalone Kokkos-based C++ programs. The agent is instructed to preserve the exact computational semantics, variable names, and function signatures while ensuring portability across all Kokkos backends such as CUDA, HIP, and OpenMP. A key requirement added to the translator prompt was the explicit use of Kokkos::fence() inside main() exactly once, after all parallel computations and loops and immediately before measuring or printing any results. This was necessary because the agent would sometimes omit it, despite the fence being essential for completing all outstanding asynchronous operations. It also provided a standardized

reference point across all kernels where functionality testing code could be injected consistently. To ensure complete implementations for larger kernels, the translator prompt was further augmented with directives such as “Implement all computational logic fully.”, since otherwise the generated outputs would sometimes only contain placeholder comments for the main computational sections. The translated program uses Kokkos constructs such as Kokkos::View for array allocations and supports command-line arguments for input size ( $n$ ) and number of kernel repetitions. Further, it also adds capability to run the computational loop ‘repetitions’ times and measure the total execution time across all the repetitions, and also printing the execution time to standard output with six decimal places.

The generated code is then evaluated by a *Validator Agent*, which verifies that the output is syntactically valid C++ and free from any non-code content such as comments, markdown, or natural language. If the code fails validation, it is passed to a *Fixer Agent*, which applies structural and syntactic corrections if needed and removes any non-code texts while preserving the computational logic. The code generated after this phase is usually wrapped inside code block tags (like ‘‘cpp and/or ‘‘), and we use a python function (clean\_code\_output) to extract the clean, compilation-ready C++ code from inside these tags using a regular expression.

2) *Compilation and Execution*: Once the code is obtained from the last phase, it is compiled via the *Build Agent*, which invokes a wrapped python-function tool (build\_kokkos\_program) to submit a SLURM job that compiles the program using architecture-specific toolchains on the desired hardware partition. This SLURM job utilizes Spack to load the correct software environment, including the appropriate version of Kokkos and backend libraries/packages. The output of the build process is monitored, and in case of

failure, the *Error Summarizer Agent* parses the stderr logs to extract the root cause and condense it into a concise plain text summary not exceeding 20 lines. If possible, it also suggests a couple potential fixes for the issues. This summary is then utilized by the *Compile Error Fixer*, which attempts to correct the C++ code based on the diagnosed issues.

Following successful compilation, the code is executed using the *Run Agent*, which invokes a wrapped python-function tool (`run_kokkos_program`) with the desired configuration and parameters and submits a runtime SLURM job on the desired hardware partition. The same Spack-managed environment is reused during execution to guarantee that runtime libraries, compiler toolchains, and backend configurations are consistent with those used at build time. GPU profiling is also enabled and is run for the maximum input size. Similar to the build phase, any runtime failures are summarized by the *Error Summarizer* and addressed by the *Runtime Error Fixer*, which patches the code based on runtime-specific errors such as segmentation faults, invalid memory access, or device synchronization issues.

3) *Functionality Testing*: After successful execution, the functional correctness of the translated program is evaluated by the *Functionality Tester* which utilizes another wrapped python-function tool (`validate_functionality`). This agent injects kernel-customized testing code into the Kokkos program, immediately after `Kokkos::fence()`, to capture the output of the relevant resultant array of the program into a CSV file. Then, this code is re-compiled to ensure it is executable after the code injection. Simultaneously, the Fortran kernel’s driver program is also compiled using GFortran with the required flags for OpenMP, etc. Once the builds are successful, the agent runs both the original Fortran kernel program and the translated Kokkos version (with GPU profiling disabled) on the same inputs and compares their outputs across a small sweep of problem sizes. For all kernels except the EP kernel, we check if the outputs mismatch by exceeding a tolerance limit. For the EP kernel, due to the pseudo-randomness associated with the results, we just check for a non-zero output instead. In case of functionality incorrectness, the generated Kokkos code and the original Fortran code is passed to the *Functionality Fixer*, which attempts to correct logic-level issues while preserving performance-portable constructs and overall structure. Each corrected version undergoes re-compilation and re-execution before re-testing. The injected code is removed from the Kokkos program once this process is finished. This current functionality testing workflow is tailored to the benchmark kernels used here and serves as a proof of concept; generalizing it to larger applications is left for future work.

4) *Optimization*: Once a functionally correct baseline is established, the pipeline enters the optimization loop. The *Optimizer Agent* is provided with profiling summaries generated via hardware GPU profilers (e.g., NVIDIA Nsight Compute or AMD ROCProfiler). For the profiling report generated by NVIDIA Nsight Compute (NCU), we parse out the suggested OPT points (e.g., Fig. 2) which highlight potential optimizations and estimated speedup for the relevant kernel loop and combine it into a summary. Unlike NCU, AMD rocprofv3 doesn’t generate a profiling report, so we capture metrics like memory access statistics, cache hit/miss rates, and occupancy metrics, and we convert this into a standard diagnostic summary based on thresholds established for improving these metrics. Based on this feedback, the agent applies structural code changes aimed at improving memory layout, loop ordering, execution policies (e.g., block sizes, vector lengths), and use of Kokkos hierarchical parallelism constructs such as `TeamPolicy` and `ThreadVectorRange`. Each optimized version is recompiled, re-executed, re-tested, and re-profiled to determine runtime and performance improvements. The same agents in the workflow used in earlier stages (Build, Run, Functionality Tester, and Fixer agents) are reused here to maintain consistency across the optimization iterations.

### C. Workflow Orchestration

The agentic pipeline is orchestrated through a modular workflow that coordinates agents sequentially across each stage. The process begins with translation of the Fortran source code into Kokkos C++, followed by validation, compilation, execution, functionality testing, and finally optimization. Each stage is modularized as an asynchronous isolated function that internally invokes one or more LLM agents and associated tools. This modular structure ensures that any stage can be extended, replaced, or reused independently without impacting the rest of the pipeline. The orchestration logic enforces fix attempt thresholds for each version run for robustness. These limits are defined by the user. If a given stage fails repeatedly (e.g., compilation, runtime, or functionality), it aborts after reaching `MAX_COMPILE_FIXES`, `MAX_RUNTIME_FIXES`, or `MAX_FUNCTIONALITY_FIXES`, respectively. Similarly, optimization rounds continue until `MAX_OPTIMIZATION_ROUNDS` is reached. All runtime artifacts, such as generated code versions, runtime measurements, etc., are versioned and stored. The agent interaction metadata and the tracing of the agentic workflow is recorded using MLflow [43]. The system also tracks token usage (input and output) for every LLM interaction, which is logged alongside runtime and profiling metrics. Code versions are stored under a version-controlled directory structure (e.g., `.v1`, `.v2`, etc.). Additionally, results such as runtimes, fix attempts, token counts, and total time elapsed is appended to a summary CSV after each version run for post-analysis. To ensure reproducibility, cleanup routines are called before and after each run, maintaining isolation across experiments. The combination of asynchronous execution, version tracking, and bounded iteration control makes the workflow capable of being fully autonomous.

<b>OPT</b> Memory is more heavily utilized than Compute: Look at the Memory Workload Analysis section to identify the DRAM bottleneck. Check memory replay (coalescing) metrics to make sure you're efficiently utilizing the bytes transferred. Also consider whether it is possible to do more work per memory access (kernel fusion) or whether there are values you can (re)compute.		
> Key Performance Indicators: Use the guidance related to these metrics to increase performance.		
Metric Name	Value	Guidance
gpu_compute_memory_throughput.avg	66.556331	66.556 - 19.458 >= 10.000
pct_of_peak_sustained_elapsed		

Fig. 2. Example NVIDIA Nsight Compute OPT suggestion

## VI. EXPERIMENTAL SETUP

The primary objective of our experimental evaluation is to benchmark the pipeline’s ability to autonomously produce correct and efficient C++ code across heterogeneous HPC architectures using different LLMs.

### A. Hardware and System Configuration

All experiments were conducted on GPU-accelerated high-performance computing (HPC) nodes equipped with either AMD or NVIDIA architectures. A summary of hardware partition specifications is provided in Table II. Kernels were compiled and executed using SLURM-managed jobs, and architecture-specific toolchains were selected using Spack-based environments. Software versions utilized in the workflow are listed in Table III.

### B. LLM Inference Setup

To benchmark different LLMs with our workflow, we employed both proprietary and open-source models as listed in Table IV without any LLM fine-tuning. Moreover, this paper focuses on demonstrating the feasibility of using an autonomous agentic workflow using state-of-the-art LLMs. Fine-tuning could be explored in future work to potentially improve performance further, but it is outside the scope of this paper. Proprietary LLMs, such as OpenAI’s models, were accessed via the official OpenAI API. Since these models are usage-based, we had to use tokens for inference and directly account for their costs, which are later reported in our results. Llama 4 Maverick (open-source) was served locally on HPC nodes via Ollama [44] which is a containerized inference engine for LLMs, and the requests to the model were routed through LiteLLM [45], a unified inference proxy providing a consistent API interface. LiteLLM facilitated transparent switching between different model endpoints without changing the underlying workflow logic, thereby simplifying experimentation and deployment. Additionally, we implemented a custom model provider abstraction integrated with the OpenAI Agents SDK. This provider encapsulated logic for dynamic routing based on a runtime configuration, enabling agents to transparently invoke either the OpenAI-hosted models or locally hosted Ollama-served models.

### C. Runtime Configuration

Each translated kernel was executed across a different number of input sizes ( $n$ ), sampled uniformly or logarithmically between a configurable minimum and maximum input size, MIN\_N and MAX\_N respectively, depending on the kernel. For each problem size, the kernel computation was executed for multiple repetitions ( $r$ ) within each run to capture stable timing measurements. The number of repetitions was varied across hardware backends, and for some kernels the number of iterations was scaled inversely with input size (*i.e.*, fewer iterations for larger  $n$ ) to try to keep total job runtimes within a 30-minute wall-clock limit for efficiency. For other kernels, a fixed number of iterations was used across all input sizes. This strategy normalized total compute time across large and

small input sizes while preserving computational efficiency. These runtime parameters were kept consistent across LLM variants to enable fair comparison. GPU profiling was enabled at MAX\_N in each run to extract performance diagnostics. Each kernel was run for 5 MAX\_OPTIMIZATION\_ROUNDS after the baseline run. Fix attempts for each version run during the agentic workflow were bounded by configurable thresholds: up to 20 MAX\_COMPILE\_FIXES, 20 MAX\_RUNTIME\_FIXES, and 10 MAX\_FUNCTIONALITY\_FIXES. Full runtime parameters for each kernel are shown in Table V.

### D. Optimization Evaluation

To compare optimization progress across versions, we calculate GFLOPS at the maximum input size for the kernels. GFLOPS is computed as:

$$GFLOPS = \frac{FLOPS(n, \hat{i}, r)}{t_{kernel} \times 10^9}, \quad (5)$$

where  $n$  is the input size,  $\hat{i}$  is the effective per-size iteration count (after scaling),  $r$  is the number of kernel repetitions in a single run, and  $t_{kernel}$  is the wall-clock runtime measured for the full loop over  $r$  repetitions. FLOPS are estimated with closed-form models per kernel:

- CG: tridiagonal structure from `makea()` with  $nnz = 3n - 2$  and  $c_{max} = 25$ ;  $FLOPS = r(2nnz + 3n + c_{max}(2nnz + 10n)) + (2nnz + 3n)$ .
- EP: polar method; base arithmetic cost uses RNG and accept rate  $\pi/4$ ;  $FLOPS = r(19 \cdot 2^{n+1} + 8 \cdot (\frac{\pi}{4})2^n)$ ; by default, we exclude transcendental costs associated with computing square roots and logs.
- FT:  $FLOPS = r(5n^3 \log_2(n^3))$ .
- MG: constant-factor stencil work;  $FLOPS = r(576n^3)$ .
- DGEMM: with  $\alpha = 1$ ,  $\beta = 2$ ;  $FLOPS = r(2n^3 + 3n^2)$ .

Per-size iteration scaling follows the run script logic: for FT/MG/EP we scale inversely with  $n$  between a configured minimum and maximum size, clamped to a minimum of 2 iterations; DGEMM uses the configured iteration count at  $n_{min}$  and 2 thereafter; CG uses a fixed iteration cap ( $c_{max}$ ). The GFLOPS for each version thus reflects (i) the kernel’s analytic FLOP model, (ii) the actual repetitions and scaled iterations used at the largest  $n$ , and (iii) the measured kernel time.

## VII. RESULTS AND DISCUSSION

All results reported reflect a single execution of the pipeline for each kernel-model-hardware configuration. The experiments were conducted in this way due to time constraints, as repeating the full workflow multiple times would be costly and time-consuming. Because LLMs are inherently non-deterministic, some variations in the number of agent invocations, optimization strategies, or the final “best” code version are to be expected across runs; however, the overall results presented here are representative of how the pipeline operates in practice and demonstrates its feasibility as a proof-of-concept for autonomous Fortran-to-Kokkos modernization.

TABLE II  
HARDWARE SPECIFICATIONS

GPU	CPU	Architecture	Total CPU Cores
AMD Instinct MI250, 64GB HBM2e	2× AMD EPYC 7763	x86_64	128
NVIDIA Grace Hopper Superchip GH200, 96GB HBM3	ARM Neoverse-V2	aaarch64	72
NVIDIA Ampere A100, 40GiB HBM2	AMD EPYC 7763	x86_64	64

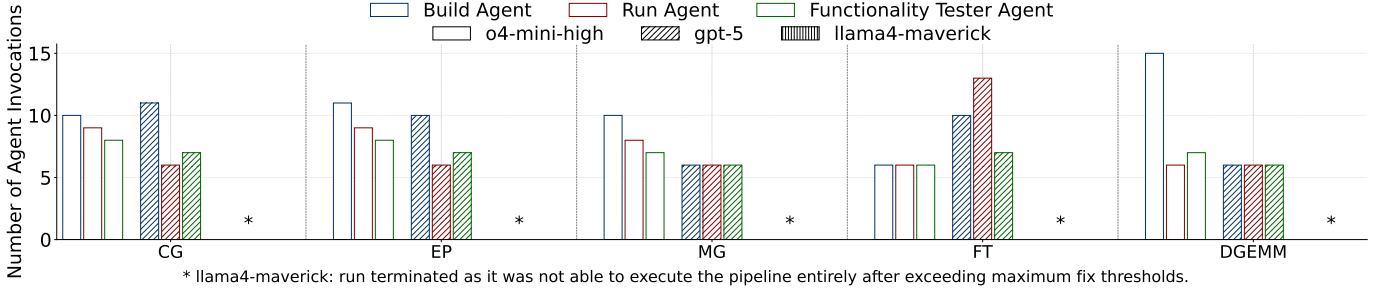


Fig. 3. Total agent invocations on AMD MI250 across benchmark kernels for the entire autonomous workflow (baseline + optimization rounds). Bars indicate the number of invocations for build, runtime, and functionality agents for different LLMs. Multiple invocations are expected since the pipeline repeatedly (i) fixes compilation and runtime errors, (ii) verifies and ensures functional correctness, and (iii) performs iterative performance optimization. Higher counts indicate more fixing cycles required before achieving a correct and optimized Kokkos implementation.

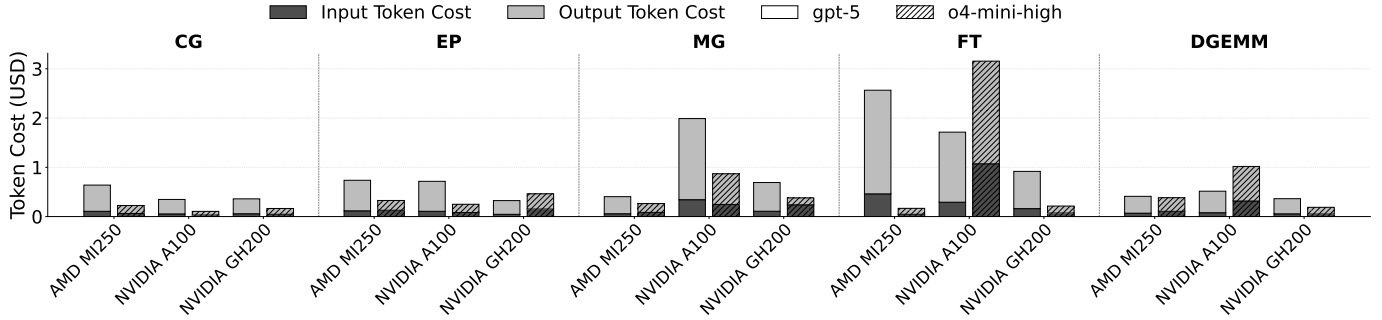


Fig. 4. OpenAI API Token costs for GPT-5 and o4-mini-high across all kernels and partitions.

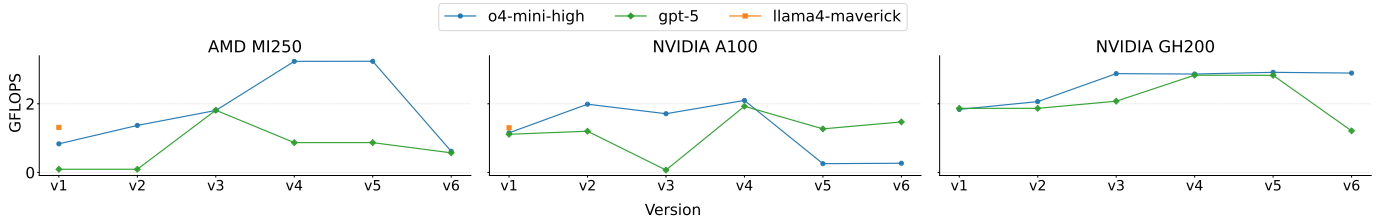


Fig. 5. Optimization trajectory of the CG kernel measured in GFLOPS at maximum input size.

TABLE III  
SOFTWARE VERSIONS

Software	Version	Software	Version
Spack	1.0.0.dev0	GCC	8.5.0
CMake	3.31.6	Kokkos	4.6.01
CUDA	12.6.3	HIP	6.4.0
Python	3.11.13	OpenAI Agents SDK	0.1.0
LiteLLM	1.73.6	Ollama	0.11.4

TABLE IV  
LLMS USED

Model	Provider	Size (parameters)	Input Cost (per 1M tokens)	Output Cost (per 1M tokens)
GPT-5 [19]	OpenAI	Unknown	\$1.25	\$10
o4-mini-high [20]	OpenAI	Unknown	\$1.10	\$4.40
Llama 4 Maverick [21]	Meta	17B active 400B total	-	-

TABLE V  
RUNTIME SETTINGS FOR DIFFERENT KERNELS AND PARTITIONS

Kernel	MIN_N	MAX_N	Num. of Sizes (n)	Program Iters.	Kernel Reps (r) (MI250, A100, GH200)
CG	1000	1000000	10	10	10, 1000, 1000
EP	18	28	5	2-5	5, 50, 50
MG	32	256	10	2-10	10, 250, 500
FT	32	128	5	2-5	10, 100, 100
DGEMM	1024	8192	5	2-5	2, 5, 5

Fig. 3 summarizes the number of agent invocations required by different LLMs to complete the pipeline on the AMD MI250. GPT-5 and o4-mini-high consistently executed all stages, though with variation in the number of build, run, and functionality tester agent invocations. Llama4-Maverick



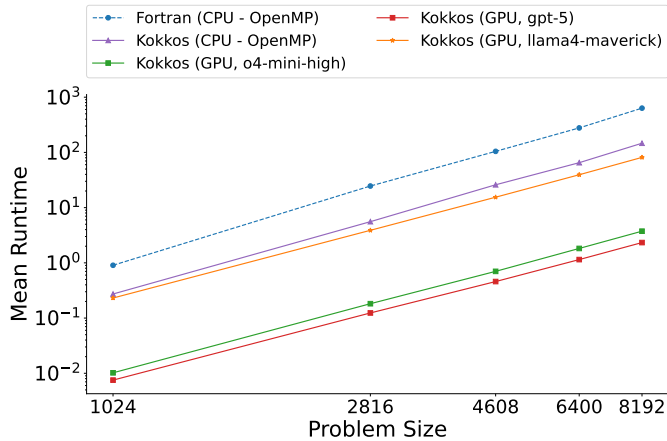


Fig. 6. Runtime comparison of DGEMM on NVIDIA A100 across codes for Fortran CPU, Kokkos CPU, and Kokkos GPU (the most optimized code version generated by the LLM).

was able to execute the baseline (v1) run for CG and reach the baseline plus a few optimization rounds for DGEMM; however, it did not succeed in completing the full workflow for any kernel by exceeding the maximum fix thresholds defined earlier. This highlights the superiority of proprietary models compared to open-source LLMs when applied to complex HPC kernels. A complete set of agent invocation results can be found in Table VII.

Fig. 4 reports the token costs incurred for OpenAI models (GPT-5 and o4-mini-high) across kernels and hardware partitions. GPT-5 incurred slightly higher costs overall, particularly due to a larger number of output tokens consumed during optimization rounds, yet both models achieved full translation and optimization of benchmark kernels for only a few U.S. dollars. This demonstrates that our autonomous approach is computationally and economically practical compared to manual modernization, which would require weeks of expert programmer time and cost.

Optimization trajectories are shown for the CG kernel in Fig. 5. Both GPT-5 and o4-mini-high demonstrated performance improvements across successive optimization versions, though it is not guaranteed that the final optimization round produced the best implementation. Occasionally, earlier versions achieved better performance than subsequent ones, underscoring the non-deterministic behavior of LLM-guided optimization. Importantly, both modules generated performance-portable code, with o4-mini-high surpassing GPT-5 on certain partitions for this specific example of the CG kernel. For Llama4-Maverick, only baseline runs were executed successfully on the AMD MI250 and NVIDIA A100, as reflected in the flat data points visible in the plot.

Fig. 6 presents runtimes for DGEMM on the NVIDIA A100 and shows how the generated codes scale with increasing problem sizes. Even on CPU backends, the Kokkos translations were faster than the original Fortran (OpenMP) version, highlighting the value of automatic modernization alone. On GPU backends, runtimes improved significantly. In this case, the most optimized code version of GPT-5 produced the fastest implementation; though o4-mini-high achieved nearly equiv-

alent performance, indicating that both models are capable of delivering high-performance optimized code. By contrast, Llama4-Maverick produced code that was significantly slower, reinforcing the observation that open-source models remain less reliable for complex HPC kernels. A potential explanation for this performance gap is that GPT-5 and o4-mini-high successfully complete multiple optimization rounds, improving memory access patterns and parallel tiling after each baseline run. In contrast, Llama4-Maverick frequently triggers the fix limits during these iterations, preventing further optimization of the generated baseline Kokkos code.

TABLE VI  
ROOFLINE ANALYSIS (DOUBLE PRECISION) FOR THE MOST OPTIMIZED GPT-5 VERSIONS OF EACH KERNEL ON NVIDIA A100.

Kernel ( $s = \text{input size}$ )	Achieved Performance (FLOPS)	Achieved Arithmetic Intensity (FLOPS/byte)	% of Roof Achieved
CG ( $s = 1000000$ )	$1.36 \times 10^{11}$	0.12	$\sim 70.3\%$
EP ( $s = 28$ )	$3.93 \times 10^{12}$	35301	$\sim 52.4\%$
MG ( $s = 256$ )	$6.99 \times 10^{11}$	0.58	$\sim 77.5\%$
FT ( $s = 128$ )	$1.48 \times 10^{11}$	1.52	$\sim 6.3\%$
DGEMM ( $s = 8192$ )	$1.88 \times 10^{12}$	15.17	$\sim 25.1\%$

*Note:* GPU peak:  $\sim 7.5 \times 10^{12}$  FLOPS, ridge point:  $\sim 4.8$  FLOPS/byte.  
% of roof achieved reports % of memory bandwidth boundary for memory-bound kernels and % of peak performance for compute-bound kernels, respectively.

The roofline analysis in Table VI provides additional perspective. Compute-bound kernels such as EP and DGEMM achieved notable fractions of the NVIDIA A100’s theoretical peak FP64 performance. In comparison, memory-bound kernels such as CG, MG, and FT achieved approximately 70.3%, 77.5%, and 6.3% of the memory bandwidth roof, respectively. Several factors may contribute to the lower performance in FT, including the difficulty of automatically restructuring data movement, challenges in exploiting cache hierarchies, and the limitations of profiler feedback for guiding memory optimizations. It is worth noting that for these kernels, sustaining very high fractions of peak memory bandwidth and performance using Kokkos is extremely difficult even for experienced programmers, underscoring the significance of the performance achieved here through an autonomous agentic AI pipeline.

Taken together, these results establish that an agentic AI workflow can autonomously translate, optimize, and deploy legacy Fortran kernels as portable Kokkos implementations. Both GPT-5 and o4-mini-high demonstrated strong reliability and performance, frequently producing results that were comparable and sometimes outperforming one another depending on the kernel. While open-source models such as Llama4-Maverick remain less capable at present, the progress demonstrated here suggests that continued advancements in model robustness and optimization strategies will further expand the scope of fully autonomous scientific code modernization.

## VIII. CONCLUSION AND FUTURE WORK

This work demonstrates that agentic AI can autonomously modernize legacy Fortran kernels into portable and performant Kokkos C++ programs. The workflow consistently produced functionally correct and optimized implementations across

TABLE VII

AGENT INVOCATIONS PER KERNEL, MODEL, AND HARDWARE PARTITION ACROSS THE PIPELINE FOR THE BASELINE VERSION (v1) AND SUCCESSIVE OPTIMIZATION VERSIONS (v2–v6). VALUES ARE REPORTED AS BUILD/RUN/FUNCTIONALITY TESTER (B/R/F) AGENT INVOCATION COUNTS. -/-/- INDICATES THAT THE PIPELINE WAS NOT ABLE TO EXECUTE SUCCESSFULLY AND EXCEEDED THE MAXIMUM FIX THRESHOLDS.

Kernel	Model	Partition	v1 (baseline) (B/R/F)	v2 (B/R/F)	v3 (B/R/F)	v4 (B/R/F)	v5 (B/R/F)	v6 (B/R/F)	Total (B/R/F)
CG	GPT-5	AMD MI250	1/1/1	1/1/1	1/1/2	3/1/1	3/1/1	2/1/1	11/6/7
		NVIDIA A100	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	6/6/6
		NVIDIA GH200	1/1/1	1/1/1	1/1/2	1/1/1	1/1/1	1/1/1	6/6/7
	o4-mini (high)	AMD MI250	1/1/1	1/1/1	1/1/1	3/4/1	1/1/1	3/1/3	10/9/8
		NVIDIA A100	1/1/2	1/1/1	1/1/2	1/1/1	2/1/1	1/1/1	7/6/8
		NVIDIA GH200	1/1/1	2/1/1	1/1/1	3/1/2	1/1/1	1/1/1	9/6/7
EP	GPT-5	AMD MI250	2/1/1	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
		NVIDIA A100	1/3/1	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
		NVIDIA GH200	1/3/1	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
	o4-mini (high)	AMD MI250	1/1/1	1/1/1	1/1/2	1/1/1	3/1/1	3/1/1	10/6/7
		NVIDIA A100	1/1/1	1/1/1	3/1/1	3/1/1	2/1/1	2/1/1	12/6/6
		NVIDIA GH200	2/1/1	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	7/6/6
MG	GPT-5	AMD MI250	1/1/1	1/1/1	2/1/1	4/4/2	1/1/1	2/1/2	11/9/8
		NVIDIA A100	1/1/1	1/1/1	1/1/1	1/1/4	4/1/1	6/1/1	14/6/9
		NVIDIA GH200	2/2/1	2/1/1	1/1/2	1/2/1	1/3/1	3/6/2	10/15/8
	o4-mini (high)	AMD MI250	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
		NVIDIA A100	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
		NVIDIA GH200	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
FT	GPT-5	AMD MI250	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	6/6/6
		NVIDIA A100	1/1/1	1/1/1	5/1/1	3/1/1	4/1/1	3/1/1	17/6/6
		NVIDIA GH200	1/1/1	1/1/1	1/1/1	1/1/1	2/1/1	3/1/1	9/1/1
	o4-mini (high)	AMD MI250	2/2/1	1/1/1	1/1/1	2/2/1	2/1/1	2/1/2	10/8/7
		NVIDIA A100	1/1/1	1/2/1	17/2/1	4/1/1	15/1/1	10/2/1	48/9/6
		NVIDIA GH200	1/2/2	1/1/1	1/1/2	1/1/2	2/1/1	1/1/1	7/7/9
DGEMM	GPT-5	AMD MI250	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
		NVIDIA A100	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
		NVIDIA GH200	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
	o4-mini (high)	AMD MI250	1/2/1	1/1/1	1/2/1	3/4/2	2/2/1	2/2/1	10/13/7
		NVIDIA A100	1/2/1	1/2/1	1/5/1	1/4/1	1/5/1	1/2/1	6/20/6
		NVIDIA GH200	1/1/1	1/1/2	1/3/1	1/2/1	1/2/1	1/1/1	6/10/7
Llama 4 Maverick	GPT-5	AMD MI250	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	6/6/6
		NVIDIA A100	1/2/1	2/2/1	2/2/1	1/2/1	9/12/1	1/11/1	16/31/6
		NVIDIA GH200	1/2/1	1/1/2	1/1/1	1/1/1	2/1/1	2/1/1	8/7/7
	o4-mini (high)	AMD MI250	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
		NVIDIA A100	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
		NVIDIA GH200	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-

diverse hardware, with kernels achieving substantial fractions of machine peak performance and memory bandwidth roof. Remarkably, these translations and optimizations were achieved in just a few hours, a task that would require expert programmers significantly more time and effort and would likely fall short of the performance achieved here. Paid OpenAI LLMs such as GPT-5 and o4-mini-high completed the full pipeline for only a few U.S. dollars, generating codes that significantly outperformed the original Fortran baselines. By contrast, open-source models like Llama4-Maverick often failed to complete the workflow, indicating that open-source LLMs require further development to achieve comparable reliability. These results establish agentic AI as a powerful and cost-effective paradigm for accelerating HPC code modernization, with the potential to transform how scientific applications are adapted to evolving supercomputing architectures.

While this workflow validates our approach, several avenues remain open for future work. First, this study was designed as a proof-of-concept demonstration of a fully autonomous agentic workflow rather than a large-scale evaluation. Accordingly, we selected representative benchmark kernels that, although compact, test different and complementary aspects of high-performance computation. However, we recognize that they do not capture the full complexity of large Fortran applica-

tions and may overlap with public code seen during model training. Demonstrating generality therefore requires applying the workflow to larger Fortran applications that have no C++ versions and contain multiple interconnected computational modules. However, such codes are often proprietary or export-controlled, making them difficult to use in a reproducible research setting, but this could be explored in future work. Second, the functionality testing in this framework was designed as a proof-of-concept as it is tailored to the specific benchmark kernels under evaluation. Although effective here, a more general and dynamic testing framework is needed for larger applications. Future work could use AI agents and LLMs to automatically generate and refine domain-specific unit tests for broader correctness coverage. Third, the current optimization is effective and incorporates profiler feedback, though continued development will further strengthen its capabilities. The optimization pipeline follows a sequential strategy, where each round builds on the previous one. Exploring alternative strategies, such as keeping only the best-performing versions at each stage, could improve consistency at the cost of additional runtime. Finally, in this work we used the same LLM across all agents to ensure consistency in benchmarking. Future versions could assign different models to specific tasks. For example, code-specialized models may

improve translation and error fixing, lightweight models could handle validation, and high-reasoning models may be best suited for optimization. Leveraging such heterogeneous LLMs represents a promising direction for advancing multi-agent workflows in code translation.

## SUPPLEMENTARY MATERIALS

The agent prompts, Fortran source codes, generated Kokkos source codes, and all result plots are available on Zenodo (<https://zenodo.org/records/17064942>).

## REFERENCES

- [1] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, Z. Liu, J. Berner, W. Wang, J. G. Powers, M. G. Duda, D. M. Barker, and X.-Y. Huang, "A description of the advanced research WRF model version 4," tech. rep., 2019.
- [2] F. Archambeau, N. Méchitoua, and M. Sakiz, "Code\_saturne: a finite volume code for the computation of turbulent incompressible flows," *International Journal on Finite Volumes*, vol. 1, 2004.
- [3] B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caffisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: the biomolecular simulation program," *J. Comput. Chem.*, vol. 30, pp. 1545–1614, July 2009.
- [4] M. P. Forum, "Mpi: A message-passing interface standard," tech. rep., USA, 1994.
- [5] L. Dagum and R. Menon, "Openmp: An industry-standard api for shared-memory programming," *IEEE Comput. Sci. Eng.*, vol. 5, p. 46–55, Jan. 1998.
- [6] A. Herten, "Many cores, many models: Gpu programming model vs. vendor compatibility overview," in *Proceedings of the SC '23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, SC-W '23, (New York, NY, USA), p. 1019–1026, Association for Computing Machinery, 2023.
- [7] C. Trott, L. Berger-Vergiat, D. Poliakoff, S. Rajamanickam, D. Lebrun-Grandie, J. Madsen, N. Al Awar, M. Gligoric, G. Shipman, and G. Womeldorff, "The kokkos ecosystem: Comprehensive performance portability for high performance computing," *Computing in Science Engineering*, vol. 23, no. 5, pp. 10–18, 2021.
- [8] NVIDIA, P. Vingelmann, and F. H. Fitzek, "Cuda," 2020.
- [9] AMD, "Hip: C++ heterogeneous-compute interface for portability."
- [10] T. K. G. Inc., "Sycl: C++ programming for heterogeneous parallel computing."
- [11] H. Kaiser *et al.*, "HPX - The C++ Standard Library for Parallelism and Concurrency," *Journal of Open Source Software*, vol. 5, no. 53, p. 2352, 2020.
- [12] M. Chen *et al.*, "Evaluating Large Language Models Trained on Code," 2021.
- [13] OpenAI, "Addendum to openai o3 and o4-mini system card: Codex."
- [14] B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, R. Sauvestre, T. Remez, J. Rapin, A. Kozhevnikov, I. Evtimov, J. Bitton, M. Bhatt, C. C. Ferrer, A. Grattafiori, W. Xiong, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, and G. Synnaeve, "Code llama: Open foundation models for code," 2024.
- [15] OpenAI, "Gpt-4 technical report," 2024.
- [16] R. A. Poldrack, T. Lu, and G. Beguš, "Ai-assisted coding: Experiments with gpt-4," 2023.
- [17] D. Bailey, E. Barszcz, J. Barton, D. Browning, R. Carter, L. Dagum, R. Fatoohi, P. Frederickson, T. Lasinski, R. Schreiber, H. Simon, V. Venkatakrishnan, and S. Weeratunga, "The nas parallel benchmarks," *Int. J. High Perform. Comput. Appl.*, vol. 5, p. 63–73, Sept. 1991.
- [18] X. Zhang, "Openblas: A new optimized blas library," in *2013 IEEE International Conference on Cluster Computing (CLUSTER)*, pp. 1–4, IEEE, 2013.
- [19] OpenAI, "GPT-5 System Card," <https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf>, 2025.
- [20] OpenAI, "OpenAI o3 and o4-mini System Card," <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>, 2025.
- [21] M. AI, "Llama 4: The next generation of multimodal intelligence," <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, Apr. 2025. Accessed: 2025-08-11.
- [22] S. I. Feldman, "A fortran to c converter," *SIGPLAN Fortran Forum*, vol. 9, p. 21–22, Oct. 1990.
- [23] D. Quinlan and C. Liao, "The ROSE source-to-source compiler infrastructure," in *Cetus users and compiler infrastructure workshop, in conjunction with PACT*, vol. 2011, p. 1, Citeseer, 2011.
- [24] C. Lattner and V. Adve, "Llvm: A compilation framework for lifelong program analysis & transformation," in *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-Directed and Runtime Optimization*, CGO '04, (USA), p. 75, IEEE Computer Society, 2004.
- [25] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun, "ChatDev: Communicative agents for software development," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand), pp. 15174–15186, Association for Computational Linguistics, Aug. 2024.
- [26] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, C. Zhang, J. Wang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, and J. Schmidhuber, "Metagpt: Meta programming for a multi-agent collaborative framework," 2024.
- [27] H. Le, H. Chen, A. Saha, A. Gokul, D. Sahoo, and S. Joty, "Codechain: Towards modular code generation through chain of self-revisions with representative sub-modules," 2024.
- [28] N. Ashrafi, S. Bouktif, and M. Mediani, "Enhancing llm code generation: A systematic evaluation of multi-agent collaboration and runtime debugging for improved accuracy, reliability, and latency," 2025.
- [29] P. Diehl, N. Nader, M. Maxim Moraru, and S. Brandt, "Llm benchmarking with llama2: Evaluating code development performance across multiple programming languages," *Journal of Machine Learning for Modeling and Computing*, 2025.
- [30] W. Godoy, P. Valero-Lara, K. Teranishi, P. Balaprakash, and J. Vetter, "Large language model evaluation for high-performance computing software development," *Concurrency and Computation: Practice and Experience*, vol. 36, Nov. 2024. Publisher Copyright: © 2024 John Wiley & Sons Ltd.
- [31] P. Valero-Lara, W. F. Godoy, K. Teranishi, P. Balaprakash, and J. S. Vetter, "Chatblas: The first ai-generated and portable blas library," in *Proceedings of the SC '24 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, SC-W '24, p. 19–24, IEEE Press, 2025.
- [32] W. Godoy, P. Valero-Lara, K. Teranishi, P. Balaprakash, and J. Vetter, "Evaluation of openai codex for hpc parallel programming models kernel generation," in *Proceedings of the 52nd International Conference on Parallel Processing Workshops, ICPP-W 2023*, p. 136–144, ACM, Aug. 2023.
- [33] D. Nichols, J. H. Davis, Z. Xie, A. Rajaram, and A. Bhatele, "Can large language models write parallel code?," in *Proceedings of the 33rd International Symposium on High-Performance Parallel and Distributed Computing*, HPDC '24, p. 281–294, ACM, June 2024.
- [34] N. R. Ranasinghe, S. M. Jones, M. Kucer, A. Biswas, D. O'Malley, A. B. Most, S. L. Wanna, and A. Sreekumar, "Llm-assisted translation of legacy fortran codes to c++: A cross-platform study," 2025.
- [35] L. Chen, B. Lei, D. Zhou, P.-H. Lin, C. Liao, C. Ding, and A. Jannesari, "Fortran2c++: Automating fortran-to-c++ translation using llms via multi-turn dialogue and dual-agent integration," 2025.
- [36] P. Valero-Lara, W. F. Godoy, J. Gonzalez, A. Huante, H. Gauthier-Chaparro, J. Gonzalez, Y. K. Tang, K. Teranishi, and J. S. Vetter, "LLM-Driven Fortran-to-C/C++ Portability for Parallel Scientific Codes," in *2025 IEEE International Conference on eScience (eScience)*, (Los Alamitos, CA, USA), pp. 385–394, IEEE Computer Society, Sept. 2025.
- [37] M. Jette, C. Dunlap, J. Garlick, and M. Grondona, "Slurm: Simple linux utility for resource management," 07 2002.
- [38] T. Gamblin, M. P. LeGendre, M. R. Collette, G. L. Lee, A. Moody, B. R. de Supinski, and W. S. Futral, "The spack package manager: Bringing order to hpc software chaos," in *Supercomputing 2015 (SC'15)*, 2015.
- [39] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang, "Autogen: Enabling next-gen LLM applications via multi-agent conversations," in *First Conference on Language Modeling*, 2024.

- [40] OpenAI, “Openai agents sdk.” <https://openai.github.io/openai-agents-python/>, 2024. Accessed: 2025-07-21.
- [41] J. D. McCalpin, “Memory bandwidth and machine balance in current high performance computers,” *IEEE Computer Society Technical Committee on Computer Architecture (TCCA) Newsletter*, pp. 19–25, Dec. 1995.
- [42] J. D. McCalpin, “Stream: Sustainable memory bandwidth in high performance computers,” tech. rep., University of Virginia, Charlottesville, Virginia, 1991-2007. A continually updated technical report. <http://www.cs.virginia.edu/stream/>.
- [43] Databricks, “MLflow: An open source platform for the machine learning lifecycle.” <https://github.com/mlflow/mlflow>, 2018. Accessed: 2025-08-05.
- [44] Ollama, “Ollama.” <https://ollama.com>. Accessed: 2025-07-21.
- [45] BerriAI, “Litellm: One api to run models across providers.” <https://github.com/BerriAI/litellm>. Accessed: 2025-07-21.