

Recent Advances in Deep Learning Kernel Optimization Using Large Language Models

TIANLIN LI*, Beihang University, China
CHENXI YANG*, Tianjin University, China
QITONG SUN, Beihang University, China
XIAOYU ZHANG, Nanyang Technological University, Singapore
QIANG HU, Tianjin University, China
ZHE TANG, Zhejiang Lab, China
SHENG CHEN, Zhejiang Lab, China
FEI YANG, Zhejiang Lab, China
AISHAN LIU, Beihang University, China
XIANGLONG LIU, Beihang University, China
CHAO SHEN, Xi'an Jiaotong University, China
YVES LE TRAON, University of Luxembourg, Luxembourg
YANG LIU, Nanyang Technological University, Singapore

The exponential growth of deep learning models, especially Large Language Models (LLMs), has dramatically increased computational demands. To meet these demands, modern deep learning systems increasingly depend on specialized hardware accelerators, such as NVIDIA GPUs. The performance of deep learning workloads ultimately depends on the efficiency of computational kernels, the fundamental operators underlying these accelerators. However, these kernels are notoriously difficult to manually generate and optimize due to complex, hardware-dependent design constraints. Recent advances in LLMs are unlocking new opportunities for automated kernel generation and optimization, offering a compelling alternative to traditional labor-intensive and expert-driven approaches. This paper presents the first comprehensive survey of deep learning kernel generation and optimization using LLMs. Moreover, we provide a systematic roadmap for improving benchmarking and generation techniques in this field.

Additional Key Words and Phrases: Large Language Models (LLMs), Deep Learning Kernels, Kernel Optimization, GPU Acceleration, Benchmarking

1 Introduction

The rapid advancement of deep learning (DL) has dramatically intensified the demand for high-performance computing [1]. To meet this demand, specialized hardware accelerators, including but not limited to NVIDIA GPUs, Huawei NPUs, and Google TPUs, have emerged as the de facto computational backbone for large-scale workloads [2–4]. These architectures, designed with massively parallel processing in mind, now power a wide spectrum of applications, from small-scale experimental models to cutting-edge large language models (LLMs) [5–7]. This massive parallel processing capability is primarily harnessed through DL kernels—programs that operate on parallel

*Equal contribution.

Authors' Contact Information: Tianlin Li, tianlin001@buaa.edu.cn, Beihang University, Beijing, China; Chenxi Yang, 3023244212@tju.edu.cn, Tianjin University, Tianjin, China; Qitong Sun, sunqt@buaa.edu.cn, Beihang University, Beijing, China; Xiaoyu Zhang, xiaoyu.zhang@ntu.edu.sg, Nanyang Technological University, Singapore, Singapore; Qiang Hu, qianghu@tju.edu.cn, Tianjin University, Tianjin, China; Zhe Tang, tangzhe@zhejianglab.org, Zhejiang Lab, Hangzhou, China; Sheng Chen, scucs@zhejianglab.org, Zhejiang Lab, Hangzhou, China; Fei Yang, yangf@zhejianglab.org, Zhejiang Lab, Hangzhou, China; Aishan Liu, liuaishan@buaa.edu.cn, Beihang University, Beijing, China; Xianglong Liu, xlliu@buaa.edu.cn, Beihang University, Beijing, China; Chao Shen, chaoshen@mail.xjtu.edu.cn, Xi'an Jiaotong University, Xi'an, China; Yves Le Traon, yves.letraon@uni.lu, University of Luxembourg, Luxembourg City, Luxembourg; Yang Liu, yangliu@ntu.edu.sg, Nanyang Technological University, Singapore, Singapore.

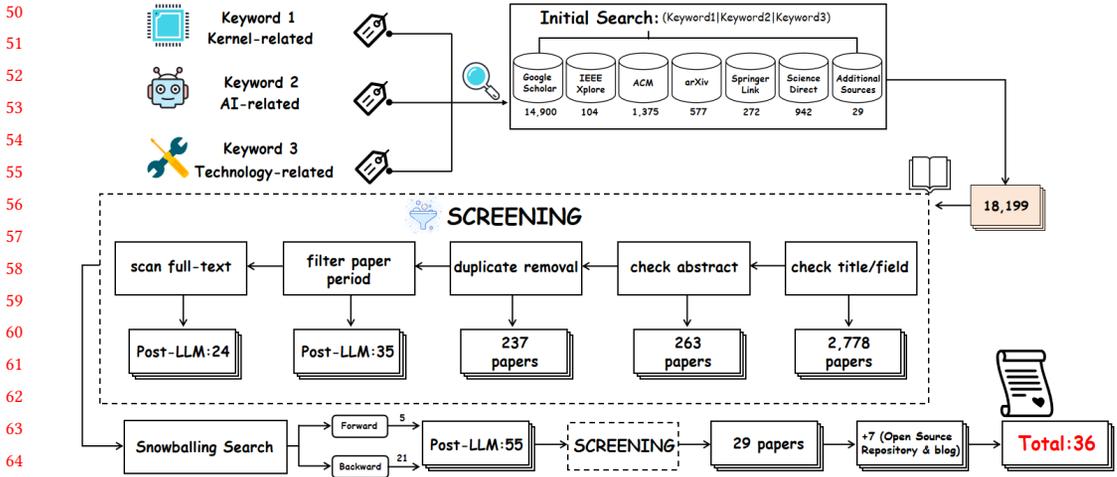


Fig. 1. The workflow and methodology of this survey.

accelerators to perform computations [8]. As the fundamental operators on these accelerators, the optimization of DL kernels is critical, directly determining the overall efficiency of the system.

Although kernels are also a type of program, their generation and optimization differ from conventional code, presenting unique and formidable challenges. The primary difficulty stems from an immense and complex optimization space defined by numerous interdependent decisions, including the configuration of thread block sizes, the intricate utilization of memory hierarchies, and the application of low-level instructions [9–11]. Furthermore, their performance is highly dependent on the specific target hardware architecture [12].

The research community has long been engaged in addressing the complex challenges of DL kernel optimization¹. Earlier approaches rely predominantly on manual tuning [13], which requires deep domain expertise to navigate a vast and highly coupled design space and to carefully balance intricate performance trade-offs among memory hierarchies, parallel execution models, and hardware-specific constraints, making manual kernel optimization both costly and difficult to scale.

Recently, the remarkable progress of LLMs has opened new opportunities for DL kernel optimization. Their strong capabilities in language understanding and code generation motivate researchers to explore how these models can assist in generating and optimizing high-performance kernels [14]. To the best of our knowledge, more than 35 LLM-driven studies have investigated kernel optimization from diverse perspectives within a single recent year, reflecting a clear and accelerating trend toward LLM-based automation.

Despite the rapid advancement and promising capabilities of LLM-based kernel generation and optimization, there is currently no systematic and comprehensive study of this emerging research area. Existing works are largely scattered across different venues and explore diverse methodological directions. This underscores the need for a timely survey of this field. To fill this gap, this survey provides a comprehensive review and comparison of existing methods. We further analyze open challenges and outline promising future research directions to guide the continued development of this field.

¹For simplicity, we do not distinguish between kernel generation and optimization, and use the two terms interchangeably.

1.1 Survey Method

The survey method adopted for preparing this paper is based on an approach widely presented in previous surveys [15]. This method includes ❶ defining the objectives of the survey, ❷ defining research questions, ❸ selecting keywords for searching, and ❹ identifying criteria for including or excluding research. These aspects are defined below.

(1) The *objectives* of this survey are defined as follows:

O1 Provide the research community with a comprehensive catalog of DL kernel optimization methods, while tracing their development over time.

O2 Discuss directions for future research to extend the research on DL kernel optimization.

(2) The *research questions* in this survey are as follows:

RQ1 How are LLM-based kernel optimization methods currently evaluated? (response in Section 3)

RQ2 How do existing approaches perform kernel optimization using LLMs? (response in Section 4)

RQ3 What roadmap can guide future research in LLM-based kernel optimization? (response in Section 5)

(3) We considered several major publication platforms, including the ACM Digital Library, IEEE Xplore, Springer, ScienceDirect, arXiv, and Google Scholar. The search strategy employed primary keywords combining *GPU/DL kernel* (or *operator*) with *generation* (or *optimization*), supplemented by additional terms such as *performance* and *system*.

(4) The resulting publications were systematically screened to identify the most relevant works. Initially, a total of 18,199 publications were considered, which were subjected to a preliminary screening based on titles and research fields, followed by an abstract-level relevance assessment. Duplicated records were then removed, and the remaining works were further filtered according to publication period to retain only post-LLM studies. A full-text screening was subsequently conducted to determine the final set of core papers. To further mitigate the risk of missing relevant studies, a snowballing search was performed on the selected core papers to identify additional candidates. These newly collected works were screened using the same multi-stage procedure, including abstract-level filtering and full-text examination. Finally, supplementary open-source repositories and technical blogs were incorporated to complement the academic literature, resulting in a curated corpus of 36 works, as illustrated in Figure 1.

The rest of this survey is organized as follows. Section 3 reviews existing performance benchmarking methodologies. Section 4 distinguishes between three major kernel optimization methodologies: single-agent systems, multi-agent systems, and training-based approaches, and organizes the surveyed methods accordingly. Section 5 outlines a forward-looking research roadmap, drawing on key insights from manual kernel optimization efforts in pre-LLM research and advances in the general code generation domain.

To support reproducibility and facilitate future research, we also provide a curated GitHub repository that catalogs all surveyed works and related resources. The repository is publicly accessible at: <https://github.com/luckily268/Awesome-GPU-Kernel-Optimization>.

2 Background and Preliminaries

Deep learning is intrinsically characterized by computationally intensive, massively parallel tensor operations, most notably large-scale matrix multiplications and convolutions. This computational paradigm is fundamentally misaligned with the architecture of general-purpose CPUs, which are designed primarily for sequential execution, complex control logic, and low-latency single-thread performance.

Table 1. Cross-vendor Mapping of GPU Architectural Terminology

Component	NVIDIA	AMD	Intel
Compute Unit	SM (Streaming Multiprocessor)	CU (Compute Unit)	Xe-core
SIMT/SIMD Group	Warp (32 threads)	Wavefront (64 threads on GCN/CDNA; 32/64 on RDNA)	Thread Group (SIMD width 8/16/32)
Scratchpad Memory	Shared Memory	LDS (Local Data Share)	SLM (Shared Local Memory)
Matrix Accelerator	Tensor (FP8/FP16/BF16/INT8)	Core MFMA (Matrix Core)	XMX (Xe Matrix Extensions)
Cache Strategy	Configurable L1/Shared partition	Vector L0 + L1 + L2 hierarchy	Configurable L1/SLM Cache/SRAM partition

To address this architectural gap, specialized hardware accelerators such as GPUs, TPUs, and NPUs have been developed. Their designs prioritize massive parallelism through thousands of streamlined cores capable of executing identical operations across large datasets with high efficiency. In the following, we provide a brief overview of the hardware design of accelerators and their operating models.

2.1 Hardware Design of Accelerators

The design of modern hardware accelerators for DL is primarily organized around two fundamental key dimensions: the *Compute Cores* and the *Memory Hierarchy*. A schematic of this organization is shown in Figure 2.

The compute cores constitute the parallel arithmetic engines. While vendor-specific terminology varies (as detailed in Table 1), such as NVIDIA’s Streaming Multiprocessors (SMs), AMD’s Compute Units (CUs), and Intel’s Xe-cores, their underlying execution philosophy converges on a throughput-first paradigm. This is realized through wide SIMT (Single-Instruction, Multiple-Threads) or SIMD (Single-Instruction, Multiple-Data) pipelines, hardware-managed multithreading, and tightly coupled on-chip storage for thread state. For instance, an NVIDIA SM executes warps of 32 threads across its SIMT pipelines; AMD CUs primarily schedule wavefronts of 64 threads; and Intel Xe-cores employ explicit SIMD vector engines with thread-group scheduling. Each such unit typically integrates scalar ALUs (for thread-private operations), wide vector ALUs, dedicated load/store pipelines, and increasingly, specialized matrix-multiply accelerators—such as NVIDIA’s Tensor Cores, AMD’s Matrix Cores (executing MFMA instructions), and Intel’s Xe Matrix Extensions (XMX). A hardware scheduler (warp/wavefront/thread-group scheduler) dynamically interleaves these thread groups to hide memory-access latency and maximize functional-unit utilization.

The memory hierarchy serves as the foundational data-supply architecture designed to meet the immense operand-throughput demands of parallel compute cores. It is structured as a multi-tiered system, spanning from high-capacity, high-bandwidth, yet high-latency off-chip memory (e.g., HBM or GDDR) down to low-latency, software-managed on-chip storage such as scratchpad memory (e.g., NVIDIA’s shared memory, AMD’s LDS, or Intel’s SLM) and dedicated L1 caches. This layered design is essential to amortize the cost of data movement, facilitate efficient data reuse, and thus determine the practical fraction of theoretical peak compute throughput that can be achieved. Consequently, the orchestration of data movement across this hierarchy, whether via cache policies, explicit DMA operations, or compiler-directed scratchpad allocation, is as critical to overall accelerator performance as the design of the computational units themselves.

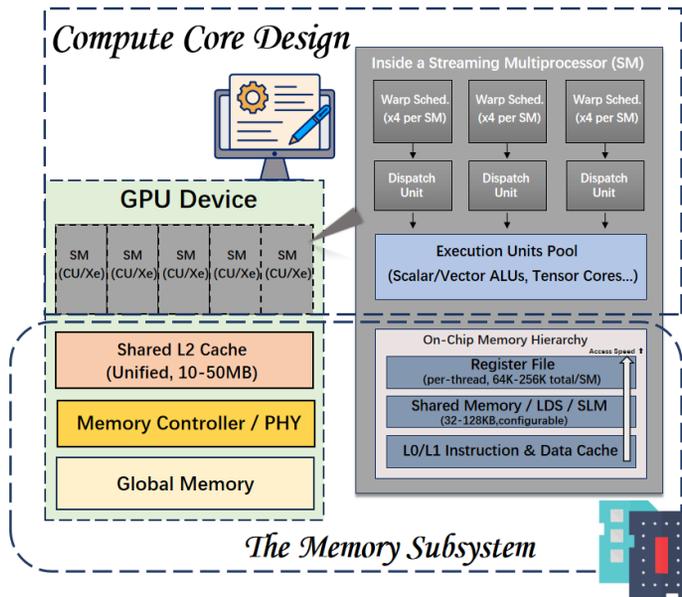


Fig. 2. GPU architecture overview. The diagram illustrates the two primary subsystems of a modern GPU. Up: Compute Core Design — The GPU comprises an array of parallel Streaming Multiprocessors (SMs, also called Compute Units or Xe-cores) that access off-chip Global Memory via a shared, unified L2 Cache and Memory Controllers/PHY. Down: Memory Subsystem within an SM — Each SM contains multiple Warp Schedulers and Dispatch Units that issue instructions to an Execution Units Pool (with scalar/vector ALUs and Tensor Cores). Data is supplied through a dedicated on-chip hierarchy: the fastest Register File, software-managed Shared Memory/LDS/SLM, and hardware-managed L0/L1 Caches.

2.2 Operating Models of Accelerators

The performance of hardware accelerators is largely influenced by the compute kernels executed on their cores. These kernels are programmed directly using architecture-specific, low-level languages and instruction sets. For NVIDIA GPUs, this involves CUDA C++ and PTX assembly; for AMD GPUs, HIP C++ and the ROCm toolchain; and for specialized units like Tensor Cores, vendor-provided intrinsic functions or microcode-level APIs. Programming at this level grants experts precise control over thread scheduling, memory access patterns, and execution pipelines, enabling the handcrafted optimization of kernels that can approach the theoretical peak throughput of the hardware.

However, developing kernels using such low-level programming remains challenging for most practitioners. To facilitate easier development, modern DL frameworks and software stacks provide developers with progressively higher-level programming interfaces for operating hardware, rather than relying on low-level programming. For better illustration, we will trace the execution path of a computational operator through the PyTorch stack on a CUDA-capable GPU. The following dissects this stack layer by layer, showing how a user's Python code is transformed, optimized, and ultimately executed as efficient, hardware-native kernels.

The typical execution path for an operator through this stack can be broken down into four stages: ① Python API Invocation, where the user's operator call is captured; ② C++ ATen Dispatching, which dynamically selects the appropriate backend kernel; ③ CUDA Kernel Launch, where execution commands are prepared for the GPU; ④ GPU Hardware Execution, where the kernel runs

on parallel streaming multiprocessors. In addition, we illustrate the Triton path as an alternative execution path extending the traditional ATen flow for comparison.

2.2.1 Python API Invocation. First, developers write Python scripts that call PyTorch operators, for example:

```

1 import torch
2
3 # Initialize tensors
4 a = torch.tensor([1.0, 2.0], device='cuda')
5 b = torch.tensor([3.0, 4.0], device='cuda')
6
7 # Element-wise addition on GPU
8 c = a + b

```

Listing 1. GPU Tensor Operations in PyTorch

When the Python interpreter executes this code, it calls PyTorch’s Python API. PyTorch’s Python API is bound to the underlying C++ implementation through PyBind11.

When calling `a + b`, it actually invokes the `__add__` method of the `torch.Tensor` object, which calls the underlying C++ function through PyBind11.

2.2.2 C++ ATen Dispatching. With the C++ function call and tensor metadata from the previous stage, PyTorch’s ATen tensor library takes over as the central dispatch engine. ATen’s role is to dynamically route each tensor operation under PyTorch’s dynamic execution model to the appropriate low-level kernel implementation, thereby decoupling the flexible, user-facing Python API from hardware-specific code. It provides a unified and extensible interface that bridges PyTorch’s runtime-dynamic graph to a wide range of backend implementations, including CPU, CUDA, XPU, and ROCm.

When an operation such as `a + b` arrives from the Python layer, ATen inspects the tensor metadata, device, data type, shape, and other attributes, and dynamically dispatches the call to the corresponding backend kernel. For example, for CUDA-resident float tensors, ATen selects the CUDA-optimized add kernel to execute the actual computation.

2.2.3 Kernel Launch. With the specific kernel function pointer, device pointers to tensor data, and calculated launch parameters from the previous stage, ATen calculates the optimal GPU thread configuration, including grid dimensions, thread-block dimensions, and other parameters, then asynchronously launches the kernel through the CUDA runtime. Specifically, `add_kernel<<< grid, block, 0, stream>>>`, the CUDA runtime places the kernel launch command and parameters into the command queue of the specified CUDA stream, and the CPU immediately returns to continue execution without waiting for GPU computation to complete.

2.2.4 GPU Hardware Execution. Once the kernel launch command reaches the GPU, its execution follows a fixed hardware pipeline: Grid and Thread-Block Allocation distributes work to SMs; Warp Scheduling issues thread groups to hide instruction latency; Memory Access loads data from the register file, shared memory, or global memory; Computation Execution performs arithmetic in CUDA cores or specialized units; and finally, Result Write-back stores outcomes to designated memory locations. This structured flow enables the massive data parallelism that defines GPU acceleration.

While the multi-layered PyTorch stack provides accessible abstractions that spare most developers from writing CUDA C++, which is a complex and error-prone process, this convenience comes at the cost of reduced flexibility for fine-grained kernel optimization. To increase flexibility while

295 maintaining low development overhead, Triton [16] was introduced as a high-level, Python-like
 296 language specifically designed for GPU kernel development. Triton provides a more intuitive and
 297 high-level programming interface, allowing developers to express complex GPU computation
 298 patterns without writing low-level CUDA code. At the same time, it offers fine-grained control over
 299 hardware resources, such as memory hierarchy, thread/block mapping, and vectorized operations,
 300 giving expert users the ability to tune kernels for maximum performance. This combination of ease
 301 of use and low-level control makes Triton a particularly promising tool for both rapid prototyping
 302 and performance-critical kernel development.

303 Specifically, practitioners can write Triton kernels as valid Python functions, which follow a
 304 Python-embedded domain-specific language (DSL) designed for GPU programming. Practitioners
 305 explicitly annotate such functions with `@triton.jit`, which registers them with the Triton runtime
 306 as parametric GPU kernel templates rather than executable Python code. Upon kernel invocation,
 307 the Python frontend extracts tensor metadata, including shapes, strides, data types, and device
 308 information, from the input tensors. This metadata is used by the Triton compiler to specialize
 309 the kernel for a specific execution configuration. During compilation, Triton lowers the annotated
 310 kernel through a multi-level intermediate representation (IR) pipeline and instantiates a target
 311 backend (e.g., CUDA or ROCm), applying a sequence of backend-specific optimization and code-
 312 generation passes. The resulting PTX code is then JIT-compiled by the GPU driver into native
 313 machine instructions (SASS), which are subsequently launched for execution on the GPU. The
 314 compilation and execution pipeline can be summarized as:

315
 316 Triton kernel $\xrightarrow{\text{Triton Compiler}}$ PTX $\xrightarrow{\text{JIT}}$ SASS $\xrightarrow{\text{Kernel Launch}}$ GPU Execution
 317

318 2.3 Efficiency Bottlenecks of Kernels in DL Software Stacks

319
 320 While frameworks such as PyTorch can automatically generate and dispatch computational kernels,
 321 these default implementations often fail to achieve peak hardware performance due to an inherent
 322 trade-off between generality and specialization. The underlying reasons are multifaceted.

323 ❶ While PyTorch’s automation successfully abstracts hardware complexity and ensures func-
 324 tional correctness, the kernels it employs, typically sourced from vendor libraries like cuBLAS
 325 or generated just-in-time, are designed as general-purpose implementations. These kernels must
 326 support a wide range of tensor shapes, data types, and hardware variants, which inherently prevents
 327 them from exploiting the full performance potential of any specific workload. For example, a general
 328 matrix multiplication kernel cannot assume fixed dimensions or memory layouts, and therefore
 329 cannot apply aggressive, shape-specific optimizations such as tailored loop tiling, unrolling, or
 330 memory-access patterns.

331 ❷ Moreover, PyTorch’s dynamic execution model defers key optimization decisions until runtime,
 332 limiting the scope for deep static analysis and compilation. This constraint stems from its core design:
 333 tensor properties such as shape, data type, and device are only known during execution, which
 334 prevents the compiler from making aggressive, irreversible optimizations upfront. For instance,
 335 while a matrix multiplication kernel generated at runtime must accommodate arbitrary tensor
 336 shapes, a pre-optimized kernel can be specialized for a fixed size, such as 256×256 . This specialization
 337 enables compile-time optimizations unavailable in the dynamic path: explicit orchestration of data
 338 movement across the memory hierarchy (global \rightarrow shared \rightarrow registers), fine-tuning of thread-block
 339 and grid dimensions to maximize occupancy and latency hiding, and precise alignment of data
 340 layouts and instruction sequences to exploit specialized hardware such as Tensor Cores or Matrix
 341 Cores. Consequently, specialized kernels, whether manually engineered or auto-tuned, can achieve
 342 substantially higher hardware utilization for their target workloads than the general-purpose
 343

kernels dispatched through PyTorch’s default stack, often translating to measurable, multi-fold speedups across training and inference pipelines.

The significant performance bottleneck of current general-purpose computational kernels, which fail to fully exploit the underlying hardware’s computational potential, has become a critical limitation in deep learning systems. This persistent gap has motivated the development of numerous kernel generation and optimization techniques.

2.4 Problem Definition

DL kernel generation and optimization aim to automatically produce high-performance compute kernels for a target tensor computation and to further refine their execution behavior through schedule- and hardware-level optimizations. Given an operator-level specification S (e.g., a PyTorch operator, an intermediate representation, or a natural-language description), the system must (i) generate a kernel implementation K in a backend language $B \in \{\text{CUDA, Triton, HIP, MLIR-based backends, } \dots\}$, and (ii) optimize its schedule, memory hierarchy usage, tiling strategy, parallelization mapping, and compilation parameters to achieve maximal runtime performance on a target hardware platform H .

This problem manifests through multiple input–output pathways, *including but not limited to*:

- natural language descriptions \rightarrow CUDA kernels,
- PyTorch operator specifications (or other high-level tensor programs) \rightarrow CUDA kernels,
- PyTorch operator specifications (or other high-level tensor programs) \rightarrow Triton kernels.

Formally, DL kernel generation and optimization seek to solve:

$$K^* = \arg \max_{\substack{K \in \mathcal{G}(S, B), \\ \theta \in \mathcal{O}(K)}} P(K, \theta, H), \text{ such that } K \models S,$$

where:

- $\mathcal{G}(S, B)$ is the space of candidate kernel implementations generated from specification S in backend B ,
- $\mathcal{O}(K)$ is the optimization space over schedules, tilings, memory layouts, unrolling factors, launch configurations, and compiler parameters,
- $P(K, \theta, H)$ denotes the performance of kernel K under optimization parameters θ on hardware H .

2.5 Requirements for Kernel Optimization Methods

Here, we summarize a set of requirements that kernel optimization methods are expected to satisfy.

- **Functional Correctness.** The kernel produced by the kernel optimization method is expected to be functionally correct.
- **Acceleration.** The kernel generated by the method should make full use of hardware accelerators to maximize computational acceleration.
- **Robustness.** The kernel generated by the method should demonstrate stable correctness and performance under edge-case or challenging settings, including extreme tensor shapes, atypical data types, and adversarial or rare input distributions.
- **Efficiency.** Kernel optimization methods should minimize kernel optimization time overhead.
- **Versatility.** Kernel optimization methods must handle a broad range of operators, such as convolution and attention, along with diverse tensor shapes and numeric data types, thereby ensuring generalization beyond narrow, handcrafted examples [17–19].
- **Cross-Architecture Portability.** Kernel optimization methods should support multiple hardware architectures and software backends with minimal modification.

Table 2. Overview of Benchmarks for LLM-based GPU Kernel Generation

Benchmark	Institution	Date	Core Task	Dataset Composition	Metrics
KernelBench [17]	Stanford	2025.02	Torch → CUDA	270 tasks: L1 (100), L2 (100), L3 (50), L4 (20)	$fast_p$
TritonBench [20]	Tianjin Univ.	2025.02	Torch → Triton	350 tasks: GitHub (184), PyTorch (166)	Pass@K, Speedup
Compute-eval [21]	NVIDIA	2025.04	NL → CUDA	128 programming problems	Pass@K
NPUEval [3]	AMD	2025.07	NL/Spec → Vectorized C++ for NPU	102 common ML operators for AMD NPU	Functional Correctness, Cycle-accurate, Vectorization Score
MultiKernelBench [18]	Nanjing Univ.	2025.07	Torch → Multi-backend	285 operators across CUDA, AscendC, Pallas	Pass@K, Compilation@K, SpeedUp $_{\alpha}$ @K
BackendBench [22]	Meta	2025.07	Torch → Triton	271 ops (correctness), 124 ops (performance)	Correctness, Speedup
robust-kbench [19]	Sakana AI	2025.09	Torch → CUDA	Tasks with multi-inits, shapes, forward/backward passes	Correctness, Speedup, Generalization

Table 3. Evaluation Dimensions Coverage Across GPU Kernel Generation Benchmarks

Benchmark	Functional Correctness	Acceleration Performance	Robustness	Efficiency	Versatility	Cross-Architecture Portability	Reproducibility
KernelBench	✓	✓			✓		✓
TritonBench	✓	✓		✓			
NPUEval	✓	✓	✓	✓		✓	
MultiKernelBench	✓	✓			✓	✓	
robust-kbench	✓	✓	✓		✓		

• **Reproducibility.** Kernel optimization methods should produce kernels whose behavior, both correctness and performance, is reproducible under well-controlled settings.

3 Evaluation

This section addresses **RQ1** by systematically reviewing how to evaluate LLM-based kernel generation methods. In this section, we first summarize existing benchmarks and metrics commonly used in the literature, as in Table 2 and Table 3. We then outline the challenges and present a roadmap.

3.1 Existing Evaluation Benchmarks

KernelBench [17] and **TritonBench** [20] initiate the design of systematic evaluation benchmarks and metrics for assessing both the functional correctness and acceleration performance of LLM-generated kernels.

KernelBench establishes a foundational evaluation benchmark for CUDA kernel optimization in DL workloads. It consists of 270 programming tasks spanning multiple difficulty levels (L1–L4) and adopts the $fast_p$ metric, which measures the proportion of generated kernels that are both functionally correct and achieve at least $p \times$ speedup over PyTorch baselines.

TritonBench focuses on evaluating kernel generation using the Triton DSL and provides a complementary benchmark suite of 350 real-world operators curated from GitHub and PyTorch

repositories. It evaluates functional correctness using Pass@K and measures runtime performance using Speedup relative to optimized baselines.

However, these initial benchmarks expose several critical limitations. ❶ They remain confined to NVIDIA’s ecosystem (CUDA or Triton), leaving other hardware platforms unexamined. ❷ They evaluate kernels only on a limited set of input shapes and configurations, offering little insight into robustness across diverse workloads and runtime variations.

To overcome these shortcomings, the community soon proposed two new evaluation benchmarks. **MultiKernelBench** [18] presented by wen et al. directly targeted the first limitation of platform specificity. It introduced the first benchmark supporting kernel generation for multiple backends: CUDA, AscendC (Huawei NPU), and Pallas (Google TPU). Its core innovation was a modular backend abstraction layer that decoupled evaluation logic from platform-specific toolchains, enabling fair comparison across diverse hardware.

Meanwhile, **NPUEval** [3] establishes a new benchmark for evaluating LLMs’ ability to generate vectorized kernel code for NPUs. More than just a dataset, it provides a complete open-source evaluation harness with cycle-accurate performance metrics.

Concurrently, the **robust-kbench** [19] addresses the second limitation. It evaluates kernel correctness across diverse settings, supports both forward and backward kernel optimization, and is designed for realistic downstream applications.

Additionally, the **BackendBench** framework [22] proposed by Meta further systematically validate the functional correctness and performance of kernels generated by LLMs in real deployment. The framework conducts comprehensive functional correctness tests on 271 operators based on TorchBench and PyTorch’s OpInfo, ensuring consistency with standard implementations. Meanwhile, using real tensor shapes from models in Huggingface, it performs performance testing on 124 commonly used operators to evaluate their execution efficiency under practical workloads. Furthermore, BackendBench introduces a *success rate across attempts* metric to assess the stability of the generation process.

3.2 Challenges in Benchmarking

Kernel generation benchmarking faces multiple challenges at both the individual kernel level and the benchmark suite level. Following the requirements outlined in subsection 2.5, we organize these challenges accordingly.

Functional Correctness. Many existing kernel benchmarks provide reference implementations or fixed input sets to evaluate correctness. While these allow basic verification, they typically cover only a limited range of shapes, data types, and input distributions. This limited coverage means that kernels may appear correct under benchmark conditions but could fail in more diverse or realistic scenarios. Addressing this limitation is challenging because designing inputs that comprehensively reflect real workloads is nontrivial.

Acceleration Performance. Existing benchmarks often measure runtime to evaluate kernel acceleration, but achieving consistent and reliable measurements is difficult due to hardware variability, warm-up effects, and differences across frameworks and backends. These factors can make it hard to determine whether observed performance improvements reflect true optimization or are influenced by external conditions. This highlights the need for standardized execution environments and well-defined measurement protocols, including warm-up runs and iteration counts, to ensure reproducible results.

Robustness. While many benchmarks focus on typical inputs, kernels must also maintain correctness and stable performance under extreme or rare edge-case scenarios, such as unusual tensor shapes, atypical data distributions, or boundary batch sizes. Existing evaluations often

491 overlook these conditions, leaving kernels vulnerable to failures or performance degradation in
492 challenging situations.

493 **Efficiency.** Existing evaluations rarely quantify the efficiency of kernel generation and optimiza-
494 tion pipelines, leaving methods that are slow or resource-intensive insufficiently assessed.

495 **Versatility.** Existing benchmarks, such as NPUEval and ComputeEval, often focus on a limited
496 set of operators, shapes, and data types. While these benchmarks provide useful snapshots of kernel
497 behavior, their narrow scope makes it difficult to draw general conclusions about performance,
498 correctness, or robustness across a full spectrum of workloads. This limited coverage presents a key
499 challenge because kernels may perform well on benchmarked operators but fail or underperform
500 on untested ones.

501 **Cross-Architecture Portability.** Many existing benchmarks are tailored to a single hardware
502 platform, which limits the ability to compare kernel performance or correctness across different
503 architectures. This focus on a single platform makes it challenging to assess whether optimizations
504 generalize beyond the target hardware, and it reduces the relevance of benchmark results for
505 broader deployment.

506 **Reproducibility.** Many existing benchmarks report results that can vary significantly depending
507 on execution environments, framework versions, or hardware configurations. This variability makes
508 it difficult to reliably compare kernel performance or correctness across experiments and over time,
509 and it can obscure the true impact of optimization techniques. Addressing this issue is challenging
510 because even minor differences in system setup or runtime conditions can affect outcomes.

511

512 3.3 Roadmap for Advancing Benchmarking

513 This roadmap outlines strategic directions to improve benchmark design.

514 For functional correctness, benchmarks should move beyond fixed or narrowly scoped input
515 configurations and adopt multi-dimensional validation across diverse tensor shapes, data types, and
516 initialization schemes. For acceleration performance, benchmarks should emphasize standardized
517 and reproducible measurement protocols. This includes clearly defined warm-up procedures,
518 iteration counts, and isolation of execution environments, so that reported speedups more accurately
519 reflect genuine optimization effects rather than artifacts of runtime variability or framework-specific
520 behavior. For robustness, optimized kernels should be evaluated under extreme or uncommon
521 conditions, such as irregular tensor shapes, in order to expose brittle optimization strategies that
522 perform well only under nominal settings.

523 For efficiency, evaluation should extend beyond the quality of the generated kernels themselves
524 to include the cost of the generation and optimization process. Benchmarks should measure factors
525 such as generation latency, compilation overhead, search or iteration budgets, and overall resource
526 consumption, providing a more realistic assessment of practical usability. For versatility, benchmarks
527 should expand the diversity of evaluated kernels by including operators drawn from a wide range
528 of model families and application domains. Leveraging real workloads and tensor shapes from
529 modern deep learning models can help ensure that benchmark results remain representative as
530 workloads continue to evolve. For cross-architecture portability, benchmarks should adopt modular
531 and backend-agnostic designs that enable evaluation across heterogeneous hardware platforms,
532 including GPUs, NPUs, and emerging accelerators. Such designs can expose architectural biases in
533 kernel generation models and encourage the development of more portable optimization strategies.
534 For reproducibility, benchmarks should provide open, well-documented evaluation harnesses and
535 reference environments, enabling results to be reliably reproduced and compared across different
536 systems, software stacks, and time periods.

537 The emergence of more comprehensive benchmarks could serve both as a testbed for existing
538 methods and as a feedback mechanism to improve future kernel generation models.

539

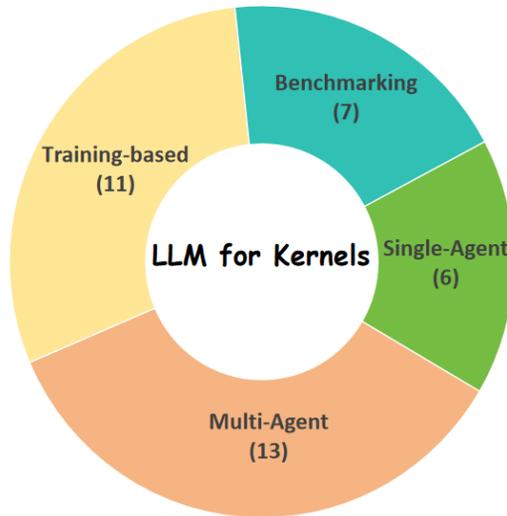


Fig. 3. Distribution of LLM-driven GPU kernel generation and optimization research. Percentages are calculated over the 37 surveyed post-LLM kernel generation and optimization papers. Works focusing primarily on mobile systems [4] are excluded from the statistics. KernelBench [17] and AI CUDA Engineer / robustkBench [19] span multiple methodological categories and are therefore counted separately. In addition, partial code releases and documentation-only resources included in Figure 1 are not considered in this statistical count.

4 Kernel Generation and Optimization Techniques

This section addresses **RQ2** by reviewing how existing approaches perform kernel generation and optimization using LLMs. Existing kernel optimization approaches can be broadly classified into three categories, as shown in Figure 3: single-agent, multi-agent, and training-based methods. This section presents a detailed overview of each category. These works are summarized in Table 4.

4.1 Single-Agent Systems

The initial wave of research on LLM-driven kernel optimization primarily centered around single-agent systems. The evolution within this category demonstrates a clear trajectory: starting from evaluating basic prompting efficacy, integrating with verifier and profiler, and culminating in reformulating kernel optimization as a structured optimization problem.

In the early exploration of automated kernel generation with LLMs, Ouyang et al. [17] introduce KernelBench and conduct a series of pilot experiments. They first adopt a one-shot prompting approach and evaluate several state-of-the-art models, including GPT-4o [52], DeepSeek-R1 [53], and Llama [54]. The results reveal that even the best models can only outperform the PyTorch baseline in fewer than 20% of the tasks under one-shot generation. While reasoning-enhanced models exhibit fewer execution errors, they still struggle with functional correctness. When tested across multiple NVIDIA GPU platforms (L40S, A100, H100, T4, etc.), the performance of generated kernels varies considerably, indicating limited model adaptability to hardware-specific characteristics. Subsequently, they also experiment with feedback-driven optimization and knowledge-augmented

Table 4. Works and Publication Dates with Open Source Status

LLM-based Code Generation Work	Date	Type	Open Source?	Benchmarking
KernelBench [17]	2025.02	Single-Agent	✓	KernelBench
Chen et al. [23]	2025.02	Single-Agent	×	KernelBench
CuAsmRL [11]	2025.03	Training-based	×	Others
Brabec et al. [24]	2025.04	Single-Agent	✓	Others
KernelLLM [25]	2025.05	Training-based	✓	KernelBench-Triton ²
CUDA-LLM [26]	2025.06	Single-Agent	×	KernelBench, CUDA Samples [27], LeetGPU [28]
GPU Kernel Scientist [29]	2025.06	Multi-Agent	×	Others
Kevin [30]	2025.07	Training-based	×	KernelBench
CUDA-L1 [31]	2025.07	Training-based	✓	KernelBench
Geak [32]	2025.07	Multi-Agent	✓	TritonBench-revised Benchmark ³
AutoTriton [33]	2025.07	Training-based	✓	KernelBench, TritonBench
Mishra and Nangia [34]	2025.07	Multi-Agent	×	Others
SwizzlePerf [12]	2025.08	Training-based	×	Others
Hao et al. [4]	2025.09	Mobile System	✓	Others
Astra [35]	2025.09	Multi-Agent	✓	Others
AI CUDA Engineer [19]	2025.09	Multi-Agent	×	robust-kbench
ConCuR [36]	2025.10	Training-based	✓	KernelBench
EVOENGINEER [37]	2025.10	Single-Agent	✓	Others
STARK [38]	2025.10	Multi-Agent	×	KernelBench
Nichols et al. [39]	2025.10	Training-based	×	KernelBench
TRITONRL [40]	2025.10	Training-based	✓	KernelBench
KernelFalcon [41]	2025.11	Multi-Agent	✓	KernelBench
CudaForge [42]	2025.11	Multi-Agent	✓	KernelBench (sampled)
PRAGMA [43]	2025.11	Multi-Agent	×	KernelBench
SparseRL [44]	2025.11	Training-based	×	Others
KERNELBAND [45]	2025.11	Single-Agent	×	TritonBench
MTMC [46]	2025.11	Training-based	×	KernelBench, TritonBench
KForge [47]	2025.11	Multi-Agent	×	Kernelbench
PIKE [48]	2025.11	Multi-Agent	×	METR-refined [49] variant of KernelBench
TritonForge [50]	2025.12	Multi-Agent	×	TritonBench
AKG [51]	2025.12	Multi-Agent	✓	KernelBench

² KernelBench-Triton is a variant of KernelBench [17], adapted specifically for evaluating Triton kernel generation.

³ TritonBench-revised is an enhanced version of TritonBench, where Wang et al. [32] corrected kernel errors and fixed missing function calls in the original evaluation suite.

Notes: ✓ = Open Source, × = Not Open Source.

Type: *Training-based, Agent, Multi-Agent, Dataset, Mobile System.*

Benchmarking: Named benchmarks are used as indicated; "Others" refers to custom or unspecified evaluation suites; "-" indicates no benchmarking was declared or applicable.

prompting. Their findings demonstrate that iterative refinement incorporating execution and manual feedback effectively helps models correct errors and discover more efficient implementations. When provided with relevant hints, models attempt to employ more advanced optimization strategies, such as shared memory or tensor core instructions, though this often increases the risk of compilation and runtime failures.

638 Following the agent design in KernelBench, Chen et al. [23] has developed a new workflow
639 that combines the DeepSeek-R1 model with verifier in a closed-loop fashion to generate opti-
640 mized attention kernels. The workflow begins with a manual prompt, and the DeepSeek-R1 model
641 generates the initial GPU kernel. The verifier, running on an NVIDIA H100 GPU, analyzes the
642 generated kernel and creates new prompts that are fed back to the model. This closed-loop approach
643 iteratively refines the code generation process and achieves 100% numerical correctness on Level-1
644 problems and 96% on Level-2 problems. These results demonstrate the potential of using advanced
645 models like DeepSeek-R1 with increased computational resources during inference to generate
646 high-performance GPU kernels.

647 To further explore how to enable LLMs to generate high-quality kernel code, Brabec et al. [24]
648 from Charles University and other institutions systematically evaluate the capability of reason-
649 ing LLMs to produce optimized CUDA code through three well-known CUDA assignments. By
650 introducing a *tutoring* mechanism (providing more detailed optimization hints and algorithmic
651 descriptions in the prompts), they find that the quality of generated code can be significantly
652 improved. For simpler CUDA tasks like computing histogram, where the optimization space is rela-
653 tively straightforward, appropriate suggestions alone enable the model to autonomously complete
654 the optimization. However, for more complex problems like k-nearest neighbors, which require in-
655 tricate parallel algorithm design, the models often fail to produce correct solutions without explicit,
656 step-by-step guidance. The study reveals that while LLMs excel at following clear instructions,
657 they struggle to make high-level optimization decisions independently when lacking adequate
658 guidance. Furthermore, the models exhibit limitations in selecting algorithmic hyperparameters,
659 underscoring the continued importance of integrating performance evaluation, or even auto-tuning
660 with LLM-based code generation.

661 **CUDA-LLM** presented by Chen et al. [26] integrates a **FSR** (Feature Search and Reinforcement)
662 framework that places the LLM in a foundational workflow (“natural language → candidate
663 generation → validation → performance optimization → prompt update”). Concretely, CUDA-
664 LLM decomposes the verifier used in prior work such as Chen et al. [23] into three separated
665 components: a *Compilation Verifier* to ensure syntactic and build correctness, a *Function Validator*
666 to check the functional correctness of the kernel, and a *Performance Profiler* to evaluate on-GPU
667 execution efficiency. This structured verifier design enables CUDA-LLM to form a feedback signal
668 over compilation validity, functional correctness, and runtime performance, thereby supporting
669 the iterative reinforcement optimization loop of FSR. However, the model itself is not trained to
670 internalize generalizable tool-usage behaviors, and directly motivates subsequent training-based
671 approaches in subsection 4.3.

672 Guo et al. [37] propose **EvoENGINEER**, a framework that abstracts LLM-based kernel opti-
673 mization into a structured evolutionary code search process. Rather than introducing another ad
674 hoc workflow, EvoENGINEER organizes code evolution into two orthogonal components: traverse
675 techniques for navigation strategies in the discrete code space and population management for
676 maintaining and selecting candidate solutions. This abstraction facilitates more effective indepen-
677 dent analysis and systematic comparison of different evolution strategies. Based on this framework,
678 they instantiate three representative variants: *EvoENGINEER-Free* that utilizes only task context,
679 *EvoEngineer-Insight* that leverages optimization insights and *EvoENGINEER-Full* that integrates
680 both historical solutions, forming a spectrum of progressively richer information integration and
681 population preservation strategies. Evaluated on 91 real-world CUDA kernels, EvoEngineer achieves
682 a principled balance between performance and correctness, with the highest averaged median
683 speedup of 2.72× over baseline CUDA kernels and a code validityrate of 69.8%, establishing a
684 principled and reusable foundation for evolutionary kernel optimization.

685
686

687 From a decision theoretic perspective, Ran et al. [45] introduce **KERNELBAND**, which reformulates kernel optimization as a hierarchical sequential decision problem under performance
688 uncertainty. Instead of treating kernel refinement as ad-hoc iteration, KERNELBAND models kernel
689 candidate selection and optimization strategy application as two coordinated bandit layers, guided
690 by profiling signals. The framework incorporates runtime behavior clustering to reduce redundant
691 exploration across similar kernels and leverages hardware profiling feedback to bias the search
692 toward promising optimization directions. Evaluated on TritonBench [20], KERNELBAND consistently
693 outperforms state-of-the-art baselines, achieving higher kernel efficiency with substantially
694 fewer tokens and exhibiting strong scalability without saturation as more computational resources
695 are available.
696

697 4.2 Multi-Agent Systems

699 Beyond single-agent paradigms, more works advance LLM-based kernel optimization by adopting
700 multi-agent architectures, where coordinated interactions among specialized LLM agents govern
701 the optimization process in place of predefined pipelines.

702 The **GPU Kernel Scientist** framework proposed by Andrews and Witteveen [29] represents an
703 early instantiation of multi-agent systems for GPU kernel optimization. It casts optimization as
704 a scientific discovery process following a *hypothesis–experiment–validation* loop, executed by a
705 fixed set of roles (Designer, Writer, Tester) under evolutionary selection. This formulation enables
706 exploration of unfamiliar or poorly documented hardware (e.g., AMD MI300) with minimal prior
707 expertise. However, the system relies on serial execution-time evaluation as its sole feedback signal,
708 lacks profiler-level guidance, and scales optimization primarily through repeated iterations, leading
709 to slow convergence.

710 Building on this paradigm, Wang et al. [32] propose **GEAK**, which introduces a redesigned agentic
711 optimization system for CUDA kernel generation. GEAK organizes kernel optimization into four
712 coordinated agent roles: *Generator*, *Evaluator*, *Reflector*, and *Optimizer*, forming a closed feedback
713 pipeline: the Evaluator performs cascaded correctness and performance checks, the Reflector
714 analyzes error traces and failures, and the Optimizer formulates targeted refinement strategies
715 that are fed back to the Generator for subsequent iterations. This fine-grained decomposition
716 enables scalable parallel exploration via inference-time compute scaling, rather than relying solely
717 on serial evolutionary iteration. Moreover, GEAK incorporates Reflexion-style feedback loops,
718 allowing failed or suboptimal kernels to be analyzed and revised through error tracing and reflective
719 reasoning. These design choices makes GEAK better suited for large kernel spaces and performance-
720 sensitive workloads. In addition, GEAK introduces AMD-focused benchmark suites (ROCm Triton
721 Benchmark), enabling rigorous cross-platform evaluation that was absent in earlier systems.

722 Mishra and Nangia [34] take a fundamentally different, search-oriented view of multi-agent
723 collaboration in “*How Many Agents to Beat PyTorch?*”. They introduce a central *Orchestrator* that
724 manages a branching search process over parallel kernel hypotheses, casting kernel optimization as
725 a structured tree search in the discrete code space. Within this orchestrated framework, a *Reasoner-
726 Agent* proposes multiple optimization strategies in natural language, which are instantiated in
727 parallel by a *Synthesis-Agent* into distinct kernel variants. Dedicated *Compile-Agent* and *Correctness-
728 Agent* aggressively prune invalid or incorrect candidates before on-GPU performance evaluation,
729 where surviving kernels compete and the winners seed the next search round. By controlling
730 branching, pruning, and termination, the *Orchestrator* prevents premature convergence and infinite
731 local refinement loops, enabling large-scale parallel exploration under inference-time compute
732 scaling. Evaluated on NVIDIA H100 GPUs, the framework achieves substantial speedups (e.g., 4.0×
733 for softmax), demonstrating that orchestrated multi-agent search can surpass both monolithic
734 agents and PyTorch baselines when sufficient compute budget is available.
735

736 Additionally, unlike prior systems that primarily generate optimized kernels from scratch, Wei
737 et al. [35] propose **Astra**, shifting the problem setting toward optimizing existing CUDA kernels
738 from SGLang [55], which is a widely deployed LLM serving framework. Astra organizes the op-
739 timization loop into four specialized agents. A *Testing Agent* constructs correctness test suites
740 and validates candidate kernels, while a *Profiling Agent* measures execution time and memory
741 behavior to provide hardware-level performance feedback. A *Planning Agent* jointly reasons over
742 correctness and profiling signals to propose targeted transformations, and a *Coding Agent* applies
743 these plans to synthesize new kernel implementations. To enable direct optimization of SGLang’s
744 highly interdependent kernels, Astra further introduces a pre-/post-processing pipeline that ex-
745 tracts kernels into stand-alone forms for optimization and subsequently reintegrates optimized
746 implementations back into the full framework for validation and benchmarking. This design allows
747 Astra to report speedups relative to the original production kernels while preserving compatibility
748 with the original framework. Evaluated on real SGLang kernels, Astra achieves consistent speedups
749 over single-agent baselines under zero-shot prompting, highlighting the practical potential of
750 multi-agent systems for maintaining and optimizing production GPU code.

751 Beyond coordinating specialized agents for code generation, testing, and profiling, the **ai cuda**
752 **engineer** framework proposed by Lange et al. [19] introduces a dedicated LLM-based verifier
753 as a central design innovation. Its key improvement lies in treating correctness verification as
754 a loop-internal, learnable optimization signal rather than a purely post-hoc execution filter. By
755 performing early “soft verification” prior to hardware execution, the verifier prunes obviously
756 incorrect candidates at the input stage rather than relying on expensive post-execution result
757 checking, thereby enabling deeper and more aggressive exploration of the kernel search space.
758 Moreover, ai cuda engineer integrates error summarization and in-context improvement into this
759 verification loop, forming a closed-loop evolutionary workflow for translating PyTorch operators
760 into optimized CUDA kernels and supporting complex transformations such as multi-operator
761 fusion.

762 Furthermore, **STARK** proposed by Dong et al. [38] advances prior LLM-based kernel optimizers
763 by redesigning kernel refinement as a tightly coordinated multi-agent process with strategic tree
764 search over persistent memory. STARK decomposes optimization into specialized planning, coding,
765 and debugging agents, and introduces grounded instructions and dynamic context windows to
766 translate high-level strategies into precise, localized CUDA code edits. Grounded instructions
767 anchor planned transformations to concrete code spans, specifying where and how to apply each
768 optimization, while dynamic context windows expose different historical attempts and feedback
769 to specific agents, enabling experience-guided planning, implementation, and debugging. This
770 design tightly couples strategic reasoning with low-level execution and balances exploration and
771 exploitation to systematically navigate the code space, mitigating common failure modes such
772 as incoherent refinements and myopic local search. Evaluated on KernelBench, STARK achieves
773 substantially higher success rates and runtime speedups (up to $10\times$ – $16\times$), particularly on kernels
774 where baseline agents struggle to produce valid implementations.

775 Meanwhile, Wang and the PyTorch Team at Meta [41] produce **KernelFalcon**, which organizes
776 kernel synthesis into a deterministic, orchestrated agent pipeline with decomposition, parallel
777 exploration, and execution-based verification. Its workflow is decomposed into specialized agents
778 responsible for operator fusion, subgraph extraction, Triton kernel synthesis, and end-to-end nu-
779 merical validation, coordinated by a central Orchestrator that manages delegation, failure handling,
780 and early-stop parallel search. Crucially, KernelFalcon adopts a verifier-first loop: candidate kernels
781 are compiled and executed against PyTorch references, and the system early-exits upon discovering
782 numerically correct implementations, enabling parallel exploration of diverse kernel realizations
783 while preserving full PyTorch semantics. KernelFalcon is the first known open agentic system to
784

785 achieve 100% correctness across all 250 L1/L2/L3 KernelBench tasks, demonstrating the effectiveness
786 of deeply orchestrated, verification-driven agent pipelines for reliable kernel synthesis.

787 In contrast to large, highly structured multi-agent frameworks, Zhang et al. [42] propose **Cud-**
788 **aForge**, a lightweight dual-agent system that separates kernel generation and evaluation into
789 a Coder–Judge loop. The Coder generates CUDA kernel candidates based on task instructions
790 and feedback from the Judge, while the Judge evaluates each candidate using correctness checks,
791 runtime profiling, and hardware metrics (e.g., GPU specifications and Nsight Compute outputs)
792 to identify bottlenecks and provide targeted optimization guidance. This iterative process allows
793 the Coder to progressively refine kernels across multiple rounds, correcting errors and improving
794 performance in a directed manner. By decoupling generation and evaluation, CudaForge achieves
795 highest correctness rate and significant performance gains over baseline approaches on Kernel-
796 bench [17] while maintaining strong practical performance. These results highlight that even a
797 minimalist agentic decomposition, when combined with iterative, hardware-aware feedback, can
798 deliver meaningful gains in real-world kernel optimization.

799 Building on lightweight, profiling agentic refinement such as CudaForge, Lei et al. [43] further
800 propose **PRAGMA**, a multi-agent framework that tightly integrates fine-grained hardware profiling
801 into the LLM optimization loop. Not only does PRAGMA rely on correctness or coarse runtime
802 feedback, but grounds iterative kernel refinement in detailed, hardware-aware performance signals
803 collected from both GPU and CPU backends. PRAGMA employs a Profiler Agent to gather low-
804 level metrics from diverse profiling tools, including Nsight Compute and Linux perf. A dedicated
805 Conductor Agent then interprets these metrics, performs bottleneck classification, and distills them
806 into high-level optimization hints. Guided by this feedback, the Coder Agent iteratively refines
807 kernel implementations, while the system explicitly preserves historically best-performing variants
808 and their profiling traces, enabling context-aware reasoning over evolving performance bottlenecks.
809 Experimental results on KernelBench [17] demonstrate that PRAGMA consistently outperforms
810 prior LLM-based approaches, achieving averaged speedups of $2.81\times$ on CPU and $2.30\times$ – $4.50\times$ on
811 GPU, and up to $10.95\times$ over baseline LLM-generated kernels. These results highlight the effectiveness
812 of reasoning based on detailed profiling feedback and explicit bottleneck interpretation.

813 Li et al. [50] propose **TritonForge**, a framework that centers on a LLM optimization pipeline for
814 Triton kernels. TritonForge incorporates specialized agents for test generation, kernel optimization,
815 and fault-aware remediation, forming a multi-stage workflow that supports automated benchmark-
816 ing, error correction, and iterative refinement. Profiling and code generation are performed in a
817 closed loop until performance converges or a predefined iteration budget is reached, enabling Tri-
818 tonForge to progressively steer Triton kernels toward high-performance implementations without
819 manual profiling expertise. Moreover, TritonForge also integrates NVIDIA Nsight Compute into
820 the optimization loop to collect low-level hardware metrics, such as memory throughput, warp
821 occupancy, and instruction stalls, and translates these profiling signals into structured feedback
822 for the LLM. Based on this feedback, the model generates targeted code modifications, including
823 changes to tiling strategies, memory layouts, and the insertion of auto-tuning directives. While this
824 profiling-guided loop enables TritonForge to progressively steer kernels toward higher performance
825 without manual profiling expertise, its iterative search exhibits limited exploration efficiency: the
826 LLM often revisits semantically similar but performance-neutral variants and tends to converge
827 prematurely to shallow performance plateaus, reflecting the lack of gradient-like guidance in
828 profiling feedback and motivating the need for stronger diversity control, adaptive stopping, and
829 memory-augmented search in future designs.

830 Furthermore, Sereda et al. [47] introduce **KForge**, which is a platform-agnostic agentic framework.
831 KForge is designed to operate across diverse accelerator backends. It combines a generation agent
832 with a performance analysis agent that interprets profiling data from heterogeneous sources,
833

834 including programmatic APIs and GUI-based tools. This work explores whether LLMs can generate
835 kernel programs for multiple hardware accelerators, leveraging both algorithmic and hardware-
836 specific optimizations. This separation between code synthesis and performance interpretation
837 enables cross-platform knowledge transfer with minimal supervision. By requiring only a single
838 example to target new hardware, KForge demonstrates that agentic optimization can generalize
839 across fundamentally different parallel programming models, such as NVIDIA CUDA and Apple
840 Metal.

841 Following this, Nagaitsev et al. [48] propose **PIKE**, a population-based multi-agent framework
842 for iterative LLM-driven kernel optimization. PIKE models optimization as a population search
843 process, where each agent corresponds to an independent LLM query and agents can be executed
844 sequentially or in parallel using the same underlying model, forming a shared verification-driven
845 evolutionary loop. The framework maintains a solution library storing the initial PyTorch model
846 and validated candidates. At each iteration, existing solutions are selected as seeds, from which new
847 kernels are generated via mutation or crossover. Candidate solutions are then compiled, functionally
848 validated, and benchmarked, optionally refined by a dedicated Error Fixing Agent (EFA), and finally
849 inserted back into the library. This loop repeats until convergence or a predefined budget is reached,
850 and can be parallelized through island-based population structures. Within this framework, PIKE
851 instantiates two representative strategies. PIKE-B (Branching Search) is an exploit-heavy, mutation-
852 only strategy that duplicates the top- k elite solutions to form each new population, rapidly refining
853 high-potential kernels under a single-island and short-term memory setting. In contrast, PIKE-O
854 (OpenEvolve-based) emphasizes exploration through crossover across multiple elite solutions and
855 island-based parallelism. Empirical results on the METR-refined variant of KernelBench [17] show
856 that exploit-heavy strategies combined with EFA achieve more effective optimization trajectories,
857 and that optimization step granularity is a key determinant of final performance.

858 Finally, industrial systems such as Huawei's **AKG** [51] framework illustrate how multi-agent
859 principles can be scaled and integrated into production compiler stacks. The AIKG subproject adopts
860 a role-specialized agent architecture, including Designer, Coder, Conductor, and Verifier agents,
861 integrated with MLIR-based compilation and retrieval-augmented generation. Unlike research
862 prototypes, AKG emphasizes extensibility, backend diversity, and workflow robustness, supporting
863 multiple hardware targets such as Ascend accelerators. Notably, although Astra, KForge all exhibit
864 forms of generalization, they generalize along fundamentally different dimensions. Astra [35]
865 attempts to extend its ability to autonomously apply a diverse set of optimization patterns across
866 different kernels, KForge [47] targets cross-platform hardware abstraction, while AKG achieves
867 ecosystem-level generalization through deep integration with compiler infrastructures. These
868 differences reflect distinct trade-offs between flexibility, control, and engineering complexity, and
869 suggest that no single notion of generalization dominates across all optimization scenarios.

870

871

4.3 Training-based Methods

872

873

874

Beyond agent collaboration, a series of works adopt supervised fine-tuning (SFT) on curated datasets
of optimized kernels, or reinforcement learning (RL) with execution-grounded rewards, enabling
models to learn common optimization patterns.

875

876

877

878

879

880

881

882

Firstly, He and Yoneki [11] propose **CuAsmRL**, which represents a form of training-based
optimization by directly operating on NVIDIA GPU SASS-level instruction schedules rather than
high-level kernel code. CuAsmRL formulates SASS scheduling as an *assembly game*, where a
reinforcement learning agent iteratively mutates instruction schedules starting from -O3-optimized
baselines and receives throughput-oriented rewards obtained through empirical GPU execution.
By learning to mimic expert-level manual scheduling behaviors, the model is able to automatically
discover superior low-level schedules. However, this extreme specialization incurs substantial

883 training cost, as reward signals must be obtained through repeated physical execution on GPUs.
884 Moreover, the lack of accurate analytical performance models for SASS-level instructions, limits
885 scalability and cross-domain generalization. Consequently, applying CuAsmRL to kernels from
886 new domains still requires domain-specific retraining and manual verification.

887 **KernelLLM** [25] curated the KernelBook dataset and employs SFT for end-to-end Triton kernel
888 generation. KernelLLM fine-tunes Llama-3.1-8B-Instruct on approximately 25,000 paired examples
889 of PyTorch modules and their corresponding Triton kernel implementations, augmented with
890 synthetically compiled samples generated via `torch.compile()` and curated code from TheStack
891 [56]. The resulting dataset, KernelBook, provides structured supervision that explicitly aligns
892 high-level PyTorch semantics with low-level Triton implementations. Trained using standard
893 instruction-based SFT, KernelLLM translates PyTorch programs into Triton kernel candidates,
894 which are validated through unit tests and `pass@k` sampling on KernelBench-Triton. Despite its
895 relatively modest parameter scale, KernelLLM achieves competitive performance with significantly
896 larger frontier models, highlighting the effectiveness of curated supervision in imparting GPU
897 programming patterns. However, as an imitation-based SFT approach, KernelLLM primarily inherits
898 the optimization strategies present in the training corpus, limiting its ability to extrapolate beyond
899 observed optimization patterns.

900 **Kevin (Kernel Devin)** [30] pioneers multi-turn reinforcement learning for CUDA kernel genera-
901 tion. For each task, Kevin samples multiple parallel trajectories, where kernels are iteratively refined
902 over several turns. Each refinement turn consists of a chain-of-thought (CoT) reasoning step and a
903 kernel generation step, where the CoT verbalizes intermediate optimization decisions, while the
904 kernel generation step concretely implements these decisions into an updated CUDA kernel, and
905 is treated as an individual training sample with execution-grounded rewards. To prevent context
906 explosion, long CoTs are discarded while compact summaries of optimization actions, together
907 with previously generated kernels and evaluation feedback, are retained to condition subsequent
908 refinement turns. Evaluated on KernelBench [17], Kevin improves kernel correctness from 56%
909 to 82% and increases mean speedup from 0.53 \times to 1.10 \times over the PyTorch Eager baseline. These
910 results demonstrate that reinforcement learning can effectively train models to reason and optimize
911 over a sequence of structured refinement steps. However, its evaluation is primarily conducted on
912 NVIDIA A100 GPUs, leaving generalization to diverse hardware architectures as an open question.

913 Following Kevin, **CUDA-L1** introduced by Li et al. [31] includes three stages: Supervised Fine-
914 Tuning with Data Augmentation, Self-Supervised Learning, and Contrastive Reinforcement Learn-
915 ing. The approach augments the training dataset with CUDA code variants generated by LLMs
916 and fine-tunes the base model on executable and correct implementations to establish founda-
917 tional CUDA knowledge. The model then iteratively generates CUDA kernels, validates their
918 correctness and executability, and trains on successfully validated examples, enabling autonomous
919 improvement without human supervision. Additionally, contrastive learning is employed with
920 execution-time rewards, training the model to distinguish between faster and slower CUDA im-
921 plementations, ultimately optimizing for superior performance. However, the CUDA-L1 approach
922 relies on iterative generation, validation, and training cycles, which makes the whole process
923 relatively time-consuming.

924 In the triton domain, Li et al. [33] introduce **AutoTriton** that represents the first dedicated RL-
925 trained model for Triton kernel synthesis, combining SFT with Group Relative Policy Optimization
926 (GRPO) reinforcement learning [57] under hybrid rewards based on rule and execution. Built on an
927 8B parameter architecture, AutoTriton first undergoes supervised fine-tuning on curated Triton
928 examples, then is further optimized using the GRPO algorithm with a hybrid reward function
929 that combines rule-based and execution-based feedback. AutoTriton demonstrates performance
930 comparable to significantly larger frontier models (e.g., Claude-3.5 Sonnet and DeepSeek-R1) across
931

932 five evaluation channels of TritonBench [20] and KernelBench [17]. The work highlights the
933 effectiveness of RL in learning high-level Triton programming patterns and hardware-specific
934 optimizations.

935 **SwizzlePerf** proposed by Tschand et al. [12] demonstrates that hardware topology-aware
936 execution mapping policies can also be internalized into model parameters through training. Instead
937 of generating full kernels, SwizzlePerf trains models to learn data-work-hardware swizzling policies
938 by modeling GPU memory hierarchy and architectural topology (e.g., AMD XCD). The learned
939 policies plan execution and storage mappings that optimize locality and cache utilization, effectively
940 transferring human hardware-software co-design knowledge into learned optimization behaviors.
941 Evaluations on ML and scientific kernels report speedups of up to $2.1\times$ and up to 70% improvements
942 in L2 cache hit rate, illustrating that end-to-end training can internalize not only code-level but
943 also hardware-mapping-level optimization strategies. However, its current scope mainly focuses
944 on cache hierarchy optimization, leaving other hardware resources under-explored.

945 **ConCuR** (Concise CUDA Reasoning) proposed by Kong et al. [36] addresses the data bottleneck
946 in LLM-driven kernel generation by introducing a data synthesis and curation pipeline. In the
947 synthesis stage, 18,162 PyTorch programs from KernelBook are expanded via parallel reasoning-
948 aware generation into 90,810 PyTorch-CoT-CUDA triplets, forming a large but noisy candidate
949 pool. In the curation stage, ConCuR jointly selects samples based on reasoning conciseness, runtime
950 speedup, and task-type balance, distilling 4,892 high-quality PyTorch-reasoning-CUDA triplets.
951 Fine-tuning QwQ-32B on ConCuR yields **KernelCoder**, improving pass@1 correctness from 18% to
952 58% on Level-1 and from 17% to 59% on Level-2, while also significantly boosting fast1 performance.

953 The framework proposed by Nichols et al. [39] trains LLM to interact with performance analysis
954 tools as part of the kernel optimization process. This approach fine-tunes models to perform tool-
955 assisted reasoning at inference time, enabling them to iteratively formulate optimization hypotheses,
956 invoke benchmarking and profiling tools, and refine kernel implementations through extended
957 reasoning chains. The training procedure employs reinforcement learning objectives based on
958 verifiable performance rewards, encouraging effective tool usage and measurable optimization
959 improvements while avoiding the need for large-scale online benchmarking during training. By
960 distilling optimization reasoning into compact models, the method amortizes performance engi-
961 neering expertise into model parameters and enables efficient deployment. Empirical evaluations
962 on GPU kernel benchmarks and real HPC applications demonstrate strong optimization capability,
963 including a reported 17% kernel-level speedup that translates into a 3% end-to-end application
964 improvement.

965 **TritonRL** proposed by Woo et al. [40] introduces an 8B-scale Triton-specialized language model
966 trained with a hierarchical and verifiable reinforcement learning pipeline designed to achieve both
967 high correctness and runtime performance while mitigating reward hacking. TritonRL combines
968 supervised fine-tuning with DeepSeek-R1 distillation and a subsequent RL stage featuring fine-
969 grained reward decomposition across correctness, efficiency, and style. Its verification framework
970 integrates enhanced rule checks with LLM judges to construct robust, verifiable rewards, enabling
971 reliable diagnosis of kernel validity and preventing reward hacking that arises from naive syntax-
972 only verification. By incorporating hierarchical reward assignment, token-level credit allocation,
973 and strategic data mixing across SFT and RL stages, TritonRL stabilizes multi-turn training and
974 yields improved kernel quality, generalization, and robustness. At the 8B scale, TritonRL surpasses
975 prior Triton-specific models including KernelLLM [25] and AutoTriton[33], demonstrating how
976 reinforcement learning can coordinate complex verification and generation workflows rather than
977 merely improving individual code quality.

981 While prior training-based approaches primarily focus on dense kernels, **SparseRL** [44] extends
982 reinforcement learning to sparsity-constrained CUDA kernel generation, where legality and per-
983 formance are tightly coupled. Unlike general kernel optimization, sparse computing introduces
984 hard structural constraints that must be respected throughout the optimization process, making
985 reward design and exploration substantially more challenging. SparseRL directly fine-tunes a lan-
986 guage model using RL to improve kernel correctness, sparsity-aware performance, and execution
987 efficiency. The method formulates sparse kernel generation as a sequential decision process, where
988 the model receives verifiable rewards based on compile success, functional correctness, sparsity
989 legality, and runtime performance. Through repeated interaction with the execution environment,
990 the model learns to apply domain specific sparse optimizations. Evaluated across a diverse suite of
991 sparse CUDA kernels, SparseRL significantly outperforms supervised baselines and demonstrates
992 strong generalization to unseen sparsity patterns. As a training-based RL approach, SparseRL
993 highlights the effectiveness of reinforcement learning in enabling models to internalize complex
994 hardware-aware optimization strategies for sparse GPU workloads.

995 Inspired by human staged optimization, Zhu et al. [46] propose the **MTMC** framework. MTMC
996 separates the complex task into two coordinated components: *Macro Thinking*, which employs RL
997 to train lightweight LLMs in efficiently exploring and learning semantic optimization strategies
998 that maximize hardware utilization, and *Micro Coding*, which leverages general-purpose LLMs to
999 incrementally implement stepwise optimization proposals. This decoupling allows the framework
1000 to navigate the vast optimization space while maintaining implementation correctness, avoiding
1001 the errors inherent in kernel generation. Evaluated on KernelBench [17], MTMC achieves near
1002 100% and 70% accuracy at Levels 1-2 and 3, over 50% than SOTA general purpose and domain
1003 finetuned LLMs, with up to 7.3x speedup over LLMs, and 2.2x over expert optimized PyTorch Eager
1004 kernels. On the TritonBench [20], MTMC attains up to 59.64% accuracy and 34x speedup.

1005 4.4 Challenges in LLM-Based Kernel Generation Methods

1006 LLM-based kernel generation faces multiple key challenges across evaluation dimensions. For
1007 functional correctness and acceleration performance, agent-based methods have achieved steady
1008 improvements by leveraging iterative feedback from compilers or runtime profiling. However, these
1009 approaches come at the cost of substantial computational overhead and multiple iterations, raising
1010 efficiency concerns. Purely training-based methods are faster but often produce semantically flawed
1011 kernels, resulting in incorrect results, degraded performance, instability under edge-case inputs, or
1012 suboptimal optimization.

1013 Robustness, versatility, cross-architecture portability, and reproducibility remain significant
1014 challenges for both agent-based and purely learning-based approaches. Models struggle to generalize
1015 across diverse operators, tensor shapes, data types, and hardware backends. The underlying cause is
1016 the scarcity of relevant training data and the models' insufficient ability to capture kernel semantics,
1017 including memory access patterns, parallel execution constraints, and data dependencies.

1018 5 Roadmap for Advancing LLM-Based Kernel Generation

1019 This section addresses **RQ3** by presenting a forward-looking roadmap for advancing LLM-based
1020 GPU kernel generation. Specifically, we identify two directions: integrating accumulated human
1021 expert knowledge and adapting ongoing technical advances in general-purpose code LLMs for
1022 kernel-specific generation. Human expertise in manual kernel optimization from the pre-LLM era
1023 encapsulates extensive domain knowledge, performance heuristics, and hardware-aware optimiza-
1024 tion strategies. Current LLMs, however, are unable to fully capture this expertise through purely
1025 data-driven training. Consequently, incorporating such human knowledge in an agentic way into
1026 the LLMs has the potential to substantially improve kernel generation. Moreover, although recent
1027
1028
1029

advances in general-purpose code models have led to substantial improvements in overall coding capabilities, these models are not specifically adapted for kernel-level optimization, leaving considerable room to adapt and extend such advances to kernel-specific generation and optimization. In the following, we accordingly organize the roadmap into two subsections.

5.1 Integrating Human Expertise

This subsection provides a systematic summary of prior human expertise that can inform performance optimization in LLMs. In particular, we emphasize key perspectives of human expertise, including mathematical equivalence transformations, data locality optimization, hardware instruction mapping, dynamic-to-static transformations, and precision–resource trade-offs. We illustrate these perspectives with a 2×2 convolution on a 3×3 input along the PyTorch-to-GPU execution pathway.

5.1.1 Mathematical Equivalence Transformations. Mathematical equivalence transformations reformulate computational problems into mathematically equivalent forms that can be implemented more efficiently.

We illustrate this principle using the im2col transformation as an example. As showed in Figure 4, with a 3×3 input and a 2×2 kernel, im2col flattens each overlapping patch into a column and concatenates them into a matrix—converting convolution from scattered dot products into a single dense matrix multiplication (GEMM). This restructuring enables the use of optimized GEMM libraries and hardware accelerators.

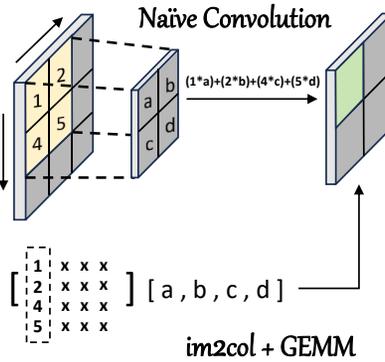


Fig. 4. Schematic comparison of native convolution versus im2col+GEMM convolution. Up: Native convolution performs a sliding-window dot product: the kernel $[a, b, c, d]$ convolves with the input patch $[1, 2, 4, 5]$ to produce one output element $(1 \cdot a) + (2 \cdot b) + (4 \cdot c) + (5 \cdot d)$. Down: im2col+GEMM first flattens all sliding windows into columns of a matrix (only the first column $[1, 2, 4, 5]^T$ is shown), then multiplies the flattened kernel $[a, b, c, d]$ with this matrix in a single GEMM operation, producing all output elements simultaneously.

5.1.2 Data Locality Optimization. Data locality optimization targets the fundamental memory bandwidth bottleneck in GPU computing. This perspective maximizes utilization of the GPU memory hierarchy from global memory through shared memory to registers by strategically restructuring data placement and access patterns to minimize data movement and maximize reuse.

1079 Specifically, coalesced memory access allows threads to load contiguous data efficiently. Tiled
1080 shared memory and kernel weight reuse in registers, combined with cache-aware layouts and bank-
1081 conflict avoidance, can also provide significant benefits [58]. Specifically, data locality optimization
1082 covers the following dimensions.

1083 **Coalesced Memory Access.** Coalescing ensures that consecutive threads within a warp access
1084 consecutive memory locations, enabling a single wide memory transaction (e.g., 128 bytes) to
1085 serve multiple threads efficiently. For regular access patterns, this can be achieved through proper
1086 data layout, thread organization, on-chip memory utilization, and techniques such as reorganizing
1087 threads [59, 60], selecting optimal thread block sizes [9], transforming data layouts (e.g., array of
1088 structs to struct of arrays), and tiling [61–63]. For irregular access patterns, such as those in sparse
1089 matrices, specialized data formats are required to maintain coalesced memory access [64].

1090 **Shared Memory Tiling.** Tiling (or spatial blocking) partitions data into blocks that fit within a
1091 streaming multiprocessor’s shared memory, enabling high-bandwidth data reuse across multiple
1092 computations. This technique is particularly effective for operations with regular access patterns
1093 such as matrix multiplication and convolution. Shared memory tiling exploits both temporal
1094 locality (reusing data across multiple operations) [65–67] and spatial locality (accessing nearby
1095 data) [68–70].

1096 **Kernel Fusion.** Kernel fusion merges multiple consecutive kernels into a single kernel, elim-
1097 inating intermediate global memory writes and reads. This optimization reduces both memory
1098 traffic and kernel launch overhead. Key benefits include improved data reuse and enhanced cache
1099 utilization [71]. However, fusion may increase register and shared memory pressure, requiring
1100 careful trade-off analysis [72].

1101 **Register Blocking.** Registers provide the fastest memory tier, with zero access latency. Register
1102 blocking (or temporal blocking) stores frequently reused values (e.g., kernel weights in convolution)
1103 in registers throughout computation; therefore this technique is especially effective for algorithms
1104 with high temporal locality [73, 74].

1105 **Prefetching.** To reduce long memory latencies, data prefetching loads data for future computa-
1106 tion steps before they are needed, overlapping memory transfers with computation. This technique
1107 is commonly applied in dense linear algebra kernels (e.g., matrix multiplication [75]) and stencil
1108 operations [66], often in combination with tiling and double buffering [76].

1109
1110 **5.1.3 Hardware Instruction Optimization.** Hardware instruction optimization focuses on mapping
1111 computation to efficient GPU instructions and scheduling them to maximize execution-unit uti-
1112 lization. At this level, the algorithmic structure and data layout are largely fixed, and performance
1113 improvements are achieved by exploiting instruction-level parallelism and hardware-specific exe-
1114 cution characteristics. Specifically, optimization must address two fundamental tensions: ❶ the
1115 severe mismatch between peak arithmetic throughput and long memory and instruction latencies,
1116 and ❷ the contention for limited on-chip resources (e.g., registers and shared memory) induced
1117 by massive thread-level parallelism. Effective instruction-level optimizations therefore require
1118 careful coordination of instruction selection, scheduling, and resource allocation to fully exploit
1119 the underlying GPU microarchitecture.

1120 Considerable human effort has historically gone into addressing the two fundamental perfor-
1121 mance tensions in GPU kernels. To mitigate the mismatch between high arithmetic throughput
1122 and long memory and instruction latencies, expert developers maximized parallelism at multiple
1123 levels. At the instruction level, they applied loop unrolling and instruction scheduling to increase
1124 instruction-level parallelism (ILP) and keep deep CUDA core pipelines occupied. Vectorization (e.g.,
1125 using float4) further enhanced throughput by enabling SIMD execution and improving memory
1126

1127

1128 coalescing [77, 78]. At the thread-group level, warp-centric programming and warp shuffle operations
 1129 facilitated efficient data exchange and reduction without shared-memory synchronization,
 1130 sustaining warp activity and improving latency hiding [79, 80].

1131 Similarly, managing contention for limited on-chip resources required careful tuning of work
 1132 granularity and execution mapping. Thread coarsening illustrated a key trade-off: assigning more
 1133 computation per thread improved ILP and register-level data reuse but increased register pressure,
 1134 potentially reducing occupancy and limiting latency hiding [81]. Likewise, offloading computation to
 1135 specialized execution units such as Tensor Cores achieved orders-of-magnitude higher throughput
 1136 for mixed-precision GEMM [82], but imposed strict constraints on data layout, problem size, and
 1137 kernel design, influencing overall resource allocation and scheduling decisions.

1138 Previous work has shown that reducing numerical precision can improve computational efficiency
 1139 and resource utilization, offering strategies that may also benefit large-scale model training and
 1140 deployment. To balance computation, memory usage, and accuracy, practitioners selectively reduced
 1141 precision, using FP16 and INT8 operations to leverage specialized hardware while maintaining
 1142 acceptable model accuracy.

1143 During training, techniques such as Automatic Mixed Precision (AMP) dynamically combine
 1144 16-bit and 32-bit operations, achieving significant speedups and memory savings without compro-
 1145 mising accuracy [83]. For deployment, aggressive post-training quantization maps weights and
 1146 activations to low-bit representations (e.g., INT4), reducing model size and enabling execution
 1147 on resource-constrained devices, while quantization-aware training mitigates accuracy loss by
 1148 embedding rounding and clipping directly into the forward pass [84]. Beyond precision alone,
 1149 co-designing data layouts with reduced-precision arithmetic—such as transforming from Array-of-
 1150 Structs to Struct-of-Arrays—has been shown to unlock additional speedups on GPUs by improving
 1151 memory access patterns and cache utilization [85].

1152 Current LLMs fall short of fully understanding human domain expertise. Integrating such
 1153 expertise in an agentic way could substantially improve LLM-driven kernel generation.

1154 5.2 Adapting Technical Advances in General-Purpose Code LLMs

1155 In this section, we distill recent advances in the general code domain into a set of high-level
 1156 principles (P1–P3) that may guide the advancement of DL kernel generation.

1157 • **P1: Execution Semantic Integration.** Most LLM-based code generation approaches predomi-
 1158 nantly operate on textual program representations, implicitly treating source code as a sufficient
 1159 proxy for execution semantics relevant to optimization. However, in general programming tasks,
 1160 it has been shown that textual representations alone are often insufficient to ensure functional
 1161 correctness.
 1162

1163 To address this limitation, some studies have explored ways to embed richer code semantics
 1164 into model training, enabling LLMs to better capture execution-level behaviors and performance-
 1165 relevant properties [86, 87]. This suggests that, for kernel generation, explicitly incorporating
 1166 semantic information such as execution-level behaviors from kernel code into LLM-based methods
 1167 could substantially enhance their ability to generate functionally correct kernels.

1168 • **P2: Performance-Aware Semantic Integration.** Performance-oriented code generation, such
 1169 as kernel optimization, differs from standard code generation in that it requires balancing runtime
 1170 efficiency and functional correctness, which could be framed as a multi-objective optimization
 1171 problem. Recent advances [88–90] have been proposed to address multi-objective optimization in
 1172 the general-purpose code domain. These developments also suggest a path forward for LLM-based
 1173 kernel generation.

1174 • **P3: Hardware-Aware Cost Modeling.** Kernel optimization is inherently hardware-aware: per-
 1175 formance depends critically on factors such as memory access patterns, cache behavior, parallelism,
 1176

and synchronization. Current LLMs, however, lack native awareness of these hardware constraints, limiting their ability to reliably generate high-performance kernels. Recent advances [91] have focused on predicting numeric outcomes of code executions, which has the potential to be applied in hardware-aware cost modeling. This could be further leveraged to improve hardware-aware kernel generation in LLMs.

6 Conclusion

This survey provides an overview of current benchmarks and techniques for LLM-driven kernel generation and optimization. Despite notable progress, systematically improving the performance of existing methods remains challenging. We summarize insights from the pre-LLM era and the broader code generation domain that may inform future advances in kernel optimization. We hope this survey serves as a timely reference and motivates further research on DL kernel generation and optimization.

References

- [1] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute trends across three eras of machine learning. In *2022 international joint conference on neural networks (IJCNN)*. IEEE, 1–8.
- [2] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. 2017. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture*. 1–12.
- [3] Sarunas Kalade and Graham Schelle. 2025. NPUEval: Optimizing NPU Kernels with LLMs and Open Source Compilers. *arXiv preprint arXiv:2507.14403* (2025).
- [4] Zixu Hao, Jianyu Wei, Tuowei Wang, Minxing Huang, Huiqiang Jiang, Shiqi Jiang, Ting Cao, and Ju Ren. 2025. Scaling LLM Test-Time Compute with Mobile NPU on Smartphones. *arXiv preprint arXiv:2509.23324* (2025).
- [5] Minh-Khoi Nguyen-Nhat, Hoang Duy Nguyen Do, Huyen Thao Le, and Thanh Tuan Dao. 2024. LLMPerf: GPU Performance Modeling meets Large Language Models. In *2024 32nd International Conference on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*. IEEE, 1–8.
- [6] Gregory Bolet, Giorgis Georgakoudis, Harshitha Menon, Konstantinos Parasyris, Niranjan Hasabnis, Hayden Estes, Kirk Cameron, and Gal Oren. 2025. Can large language models predict parallel code performance?. In *Proceedings of the 34th International Symposium on High-Performance Parallel and Distributed Computing*. 1–6.
- [7] Christoforos Kachris. 2025. A survey on hardware accelerators for large language models. *Applied Sciences* 15, 2 (2025), 586.
- [8] Sunpyo Hong and Hyesoon Kim. 2009. An analytical model for a GPU architecture with memory-level and thread-level parallelism awareness. In *Proceedings of the 36th annual international symposium on Computer architecture*. 152–163.
- [9] Arno Leist, Daniel P Playne, and Kenneth A Hawick. 2009. Exploiting graphical processing units for data-parallel scientific applications. *Concurrency and Computation: Practice and Experience* 21, 18 (2009), 2400–2437.
- [10] Benjamin F Spector, Simran Arora, Aaryan Singhal, Daniel Y Fu, and Christopher Ré. 2024. Thunderkittens: Simple, fast, and adorable ai kernels. *arXiv preprint arXiv:2410.20399* (2024).
- [11] Guoliang He and Eiko Yoneki. 2025. CuAsmRL: Optimizing GPU SASS Schedules via Deep Reinforcement Learning. In *Proceedings of the 23rd ACM/IEEE International Symposium on Code Generation and Optimization*. 493–506.
- [12] Arya Tschand, Muhammad Awad, Ryan Swann, Kesavan Ramakrishnan, Jeffrey Ma, Keith Lowery, Ganesh Dasika, and Vijay Janapa Reddi. 2025. Swizzleperf: Hardware-aware llms for gpu kernel performance optimization. *arXiv preprint arXiv:2508.20258* (2025).
- [13] Pieter Hijma, Stijn Heldens, Alessio Scocco, Ben Van Werkhoven, and Henri E Bal. 2023. Optimization techniques for GPU programming. *Comput. Surveys* 55, 11 (2023), 1–81.
- [14] Alfred Santa Molison, Marcia Moraes, Glaucia Melo, Fabio Santos, and Wesley KG Assuncao. 2025. Is llm-generated code more maintainable & reliable than human-written code. *arXiv preprint arXiv:2508.00700* (2025).
- [15] Blesson Varghese, Nan Wang, David Bernbach, Cheol-Ho Hong, Eyal De Lara, Weisong Shi, and Christopher Stewart. 2021. A survey on edge performance benchmarking. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–33.
- [16] Triton Development Community. 2026. Triton Documentation. <https://triton-lang.org/main/index.html>. Accessed: 2026-01.
- [17] Anne Ouyang, Simon Guo, Simran Arora, Alex L Zhang, William Hu, Christopher Ré, and Azalia Mirhoseini. 2025. Kernelbench: Can llms write efficient gpu kernels? *arXiv preprint arXiv:2502.10517* (2025).

- 1226 [18] Zhongzhen Wen, Yinghui Zhang, Zhong Li, Zhongxin Liu, Linna Xie, and Tian Zhang. 2025. MultiKernelBench: A
 1227 Multi-Platform Benchmark for Kernel Generation. *arXiv e-prints* (2025), arXiv-2507.
- 1228 [19] Robert Tjarko Lange, Qi Sun, Aaditya Prasad, Maxence Faldor, Yujin Tang, and David Ha. 2025. Towards robust agentic
 1229 cuda kernel benchmarking, verification, and optimization. *arXiv preprint arXiv:2509.14279* (2025).
- 1230 [20] Jianling Li, Shangzhan Li, Zhenye Gao, Qi Shi, Yuxuan Li, Zefan Wang, Jiacheng Huang, WangHaojie WangHaojie,
 1231 Jianrong Wang, Xu Han, et al. 2025. Tritonbench: Benchmarking large language model capabilities for generating
 1232 triton operators. In *Findings of the Association for Computational Linguistics: ACL 2025*. 23053–23066.
- 1233 [21] NVIDIA. 2025. compute-eval. [Online]. Available: <https://github.com/NVIDIA/compute-eval>. Accessed: Dec. 21, 2025.
- 1234 [22] Meta. 2025. BackendBench. [Online]. Available: <https://github.com/meta-pytorch/BackendBench>. Accessed: Dec. 21,
 1235 2025.
- 1236 [23] Terry Chen, Bing Xu, and Kirthi Devleker. 2025. Automating GPU Kernel Generation with DeepSeek-R1 and Inference
 1237 Time Scaling. Online: NVIDIA Blog, [https://developer.nvidia.com/blog/automating-gpu-kernel-generation-with-
 1238 deepseek-r1-and-inference-time-scaling/](https://developer.nvidia.com/blog/automating-gpu-kernel-generation-with-deepseek-r1-and-inference-time-scaling/). Accessed: Nov. 11, 2025.
- 1239 [24] Matyáš Brabec, Jiří Klepl, Michal Töpfer, and Martin Kruliš. 2025. Tutoring LLM into a Better CUDA Optimizer. In
 1240 *European Conference on Parallel Processing*. Springer, 250–263.
- 1241 [25] Zacharias V. Fisches, Sahan Paliskara, Simon Guo, Alex Zhang, Joe Spisak, Chris Cummins, Hugh Leather, Gabriel
 1242 Synnaeve, Joe Isaacson, Aram Markosyan, and Mark Saroufim. 2025. KernelLLM: Making Kernel Development More
 1243 Accessible. [Online]. Available: <https://huggingface.co/facebook/KernelLLM>. Accessed: Nov. 11, 2025.
- 1244 [26] Wentao Chen, Jiace Zhu, Qi Fan, Yehan Ma, and An Zou. 2025. CUDA-LLM: LLMs Can Write Efficient CUDA Kernels.
 1245 *arXiv preprint arXiv:2506.09092* (2025).
- 1246 [27] NVIDIA Corporation. 2025. Cuda code samples. [Online]. Available: <https://github.com/NVIDIA/cuda-samples>.
 1247 Accessed: Dec. 18, 2025.
- 1248 [28] LeetGPU. 2025. Challenges. [Online]. Available: <https://leetgpu.com/challenges>. Accessed: Dec. 18, 2025.
- 1249 [29] Martin Andrews and Sam Witteveen. 2025. GPU Kernel Scientist: An LLM-Driven Framework for Iterative Kernel
 1250 Optimization. *arXiv preprint arXiv:2506.20807* (2025).
- 1251 [30] Carlo Baronio, Pietro Marsella, Ben Pan, Simon Guo, and Silas Alberti. 2025. Kevin: Multi-turn rl for generating cuda
 1252 kernels. *arXiv preprint arXiv:2507.11948* (2025).
- 1253 [31] Xiaoya Li, Xiaofei Sun, Albert Wang, Jiwei Li, and Chris Shum. 2025. Cuda-11: Improving cuda optimization via
 1254 contrastive reinforcement learning. *arXiv preprint arXiv:2507.14111* (2025).
- 1255 [32] Jianghui Wang, Vinay Joshi, Saptarshi Majumder, Xu Chao, Bin Ding, Ziqiong Liu, Pratik Prabhanjan Brahma, Dong
 1256 Li, Zicheng Liu, and Emad Barsoum. 2025. Geak: Introducing triton kernel ai agent & evaluation benchmarks. *arXiv
 1257 preprint arXiv:2507.23194* (2025).
- 1258 [33] Shangzhan Li, Zefan Wang, Ye He, Yuxuan Li, Qi Shi, Jianling Li, Yonggang Hu, Wanxiang Che, Xu Han, Zhiyuan
 1259 Liu, et al. 2025. Autotriton: Automatic triton programming with reinforcement learning in llms. *arXiv preprint
 1260 arXiv:2507.05687* (2025).
- 1261 [34] Shikhar Mishra and Ayush Nangia. 2025. How Many Agents Does it Take to Beat PyTorch? (Surprisingly Not That
 1262 Much). <https://letters.lossfunk.com/p/how-many-agents-does-it-take-to-beat>. Accessed: 2025-11-12.
- 1263 [35] Anjiang Wei, Tianran Sun, Yogesh Seenichamy, Hang Song, Anne Ouyang, Azalia Mirhoseini, Ke Wang, and Alex
 1264 Aiken. 2025. Astra: A multi-agent system for gpu kernel performance optimization. *arXiv preprint arXiv:2509.07506*
 1265 (2025).
- 1266 [36] Lingcheng Kong, Jiateng Wei, Hanzhang Shen, and Huan Wang. 2025. ConCuR: Conciseness Makes State-of-the-Art
 1267 Kernel Generation. *arXiv preprint arXiv:2510.07356* (2025).
- 1268 [37] Ping Guo, Chenyu Zhu, Siyuan Chen, Fei Liu, Xi Lin, Zhichao Lu, and Qingfu Zhang. 2025. EvoEngineer: Mastering
 1269 Automated CUDA Kernel Code Evolution with Large Language Models. *arXiv preprint arXiv:2510.03760* (2025).
- 1270 [38] Juncheng Dong, Yang Yang, Tao Liu, Yang Wang, Feng Qi, Vahid Tarokh, Kaushik Rangadurai, and Shuang Yang. 2025.
 1271 STARK: Strategic Team of Agents for Refining Kernels. *arXiv preprint arXiv:2510.16996* (2025).
- 1272 [39] Daniel Nichols, Konstantinos Parasyris, Charles Jekel, Abhinav Bhatel, and Harshitha Menon. 2025. Integrating
 1273 Performance Tools in Model Reasoning for GPU Kernel Optimization. *arXiv preprint arXiv:2510.17158* (2025).
- 1274 [40] Jiin Woo, Shaowei Zhu, Allen Nie, Zhen Jia, Yida Wang, and Youngsuk Park. 2025. TritonRL: Training LLMs to Think
 and Code Triton Without Cheating. *arXiv preprint arXiv:2510.17891* (2025).
- [41] Laura Wang and the PyTorch Team at Meta. 2025. KernelFalcon: Autonomous GPU Kernel Generation via Deep Agents.
 [Online]. Available: <https://pytorch.org/blog/kernelfalcon-autonomous-gpu-kernel-generation-via-deep-agents/>.
 Accessed: Nov. 11, 2025.
- [42] Zijian Zhang, Rong Wang, Shiyang Li, Yuebo Luo, Mingyi Hong, and Caiwen Ding. 2025. CudaForge: An Agent
 Framework with Hardware Feedback for CUDA Kernel Optimization. *arXiv preprint arXiv:2511.01884* (2025).
- [43] Kelun Lei, Hailong Yang, Huaitao Zhang, Xin You, Kaige Zhang, Zhongzhi Luan, Yi Liu, and Depei Qian. 2025. PRAGMA:
 A Profiling-Reasoned Multi-Agent Framework for Automatic Kernel Optimization. *arXiv preprint arXiv:2511.06345*

- (2025).
- [44] Anonymous. 2025. Mastering Sparse CUDA Generation through Pretrained Models and Deep Reinforcement Learning. In *Submitted to The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=VdLEaGPYWT> under review.
- [45] Dezhi Ran, Shuxiao Xie, Mingfang Ji, Ziyue Hua, Mengzhou Wu, Yuan Cao, Yuzhe Guo, Yu Hao, Linyi Li, Yitao Hu, et al. 2025. KernelBand: Boosting LLM-based Kernel Optimization with a Hierarchical and Hardware-aware Multi-armed Bandit. *arXiv preprint arXiv:2511.18868* (2025).
- [46] Xinguo Zhu, Shaohui Peng, Jiaming Guo, Yunji Chen, Qi Guo, Yuanbo Wen, Hang Qin, Ruizhi Chen, Qirui Zhou, Ke Gao, et al. 2025. QiMeng-Kernel: Macro-Thinking Micro-Coding Paradigm for LLM-Based High-Performance GPU Kernel Generation. *arXiv preprint arXiv:2511.20100* (2025).
- [47] Taras Sereda, Tom St John, Burak Bartan, Natalie Serrino, Sachin Katti, and Zain Asgar. 2025. KForge: Program Synthesis for Diverse AI Hardware Accelerators. *arXiv preprint arXiv:2511.13274* (2025).
- [48] Kirill Nagaitsev, Luka Grbcic, Samuel Williams, and Costin Iancu. 2025. Optimizing PyTorch Inference with LLM-Based Multi-Agent Systems. *arXiv preprint arXiv:2511.16964* (2025).
- [49] METR. 2025. Measuring Automated Kernel Engineering. [Online]. Available: <https://metr.org/blog/2025-02-14-measuring-automated-kernel-engineering/>. Accessed: Dec. 18, 2025.
- [50] Haonan Li, Keyu Man, Partha Kanuparth, Hanning Chen, Wei Sun, Sreen Tallam, Chenguang Zhu, Kevin Zhu, and Zhiyun Qian. 2025. TritonForge: Profiling-Guided Framework for Automated Triton Kernel Optimization. *arXiv preprint arXiv:2512.09196* (2025).
- [51] MindSpore. 2025. Auto Kernel Generator (AKG). [Online]. Available: <https://atomgit.com/mindspore/akg>. Accessed: Dec. 18, 2025.
- [52] OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] <https://arxiv.org/abs/2410.21276>
- [53] Deepseek-AI. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 8081 (Sept. 2025), 633–638. doi:10.1038/s41586-025-09422-z
- [54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] <https://arxiv.org/abs/2302.13971>
- [55] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. 2024. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems* 37 (2024), 62557–62583.
- [56] Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. 2022. The stack: 3 tb of permissively licensed source code. *arXiv preprint arXiv:2211.15533* (2022).
- [57] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).
- [58] Pieter Hijma, Stijn Heldens, Alessio Sclocco, Ben Van Werkhoven, and Henri E Bal. 2023. Optimization techniques for GPU programming. *Comput. Surveys* 55, 11 (2023), 1–81.
- [59] Juan Gómez-Luna, José María González-Linares, José Ignacio Benavides, and Nicolás Guil. 2013. An optimized approach to histogram computation on GPU. *Machine Vision and Applications* 24, 5 (2013), 899–908.
- [60] Yang Hu, Hang Liu, and H Howie Huang. 2018. Tricore: Parallel triangle counting on gpus. In *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 171–182.
- [61] Rajib Nath, Stanimire Tomov, Tingxing" Tim" Dong, and Jack Dongarra. 2011. Optimizing symmetric dense matrix-vector multiplication on GPUs. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–10.
- [62] Wai Teng Tang, Wen Jun Tan, Rajarshi Ray, Yi Wen Wong, Weiguang Chen, Shyh-hao Kuo, Rick Siow Mong Goh, Stephen John Turner, and Weng-Fai Wong. 2013. Accelerating sparse matrix-vector multiplication on GPUs using bit-representation-optimized schemes. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. 1–12.
- [63] Rajib Nath, Stanimire Tomov, and Jack Dongarra. 2010. An improved magma gemm for fermi graphics processing units. *The International Journal of High Performance Computing Applications* 24, 4 (2010), 511–515.
- [64] Jianlong Zhong and Bingsheng He. 2013. Medusa: Simplified graph processing on GPUs. *IEEE Transactions on Parallel and Distributed Systems* 25, 6 (2013), 1543–1552.
- [65] Ahmad Abdelfattah, Azzam Haidar, Stanimire Tomov, and Jack Dongarra. 2016. Performance tuning and optimization techniques of fixed and variable size batched Cholesky factorization on GPUs. *Procedia Computer Science* 80 (2016), 119–130.

- [66] Justin Holewinski, Louis-Noël Pouchet, and Ponnuswamy Sadayappan. 2012. High-performance code generation for stencil computations on GPU architectures. In *Proceedings of the 26th ACM international conference on Supercomputing*. 311–320.
- [67] Youngdong Do, Hyungmo Kim, Pyeongseok Oh, Daeyoung Park, and Jaejin Lee. 2019. SNU-NPB 2019: parallelizing and optimizing NPB in OpenCL and CUDA for modern GPUs. In *2019 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 93–105.
- [68] Minquan Fang, Jianbin Fang, Weimin Zhang, Haifang Zhou, Jianxing Liao, and Yuangang Wang. 2018. Benchmarking the GPU memory at the warp level. *Parallel Comput.* 71 (2018), 23–41.
- [69] Filip Petrovič, David Střelák, Jana Hozzová, Jaroslav Ol’ha, Richard Trembecký, Siegfried Benkner, and Jiří Filipovič. 2020. A benchmark set of highly-efficient CUDA and OpenCL kernels and its dynamic autotuning with Kernel Tuning Toolkit. *Future Generation Computer Systems* 108 (2020), 161–177.
- [70] Jianyu Huang, Chenhan D Yu, and Robert A van de Geijn. 2020. Strassen’s algorithm reloaded on GPUs. *ACM Transactions on Mathematical Software (TOMS)* 46, 1 (2020), 1–22.
- [71] Matthias Korch and Tim Werner. 2018. Accelerating explicit ODE methods on GPUs by kernel fusion. *Concurrency and Computation: Practice and Experience* 30, 18 (2018), e4470.
- [72] Jesús Carabaño, Jan Westerholm, and Tapani Sarjakoski. 2018. A compiler approach to map algebra: automatic parallelization, locality optimization, and GPU acceleration of raster spatial analysis. *GeoInformatica* 22, 2 (2018), 211–235.
- [73] Nhat-Phuong Tran, Myungho Lee, and Dong Hoon Choi. 2015. Memory-efficient parallelization of 3D lattice Boltzmann flow solver on a GPU. In *2015 IEEE 22nd International Conference on High Performance Computing (HiPC)*. IEEE, 315–324.
- [74] Weidong Qiu, Zheng Gong, Yidong Guo, Bozhong Liu, Xiaoming Tang, and Yuheng Yuan. 2016. GPU-Based High Performance Password Recovery Technique for Hash Functions. *J. Inf. Sci. Eng.* 32, 1 (2016), 97–112.
- [75] Pham Nguyen Quang Anh, Rui Fan, and Yonggang Wen. 2016. Balanced hashing and efficient gpu sparse general matrix-matrix multiplication. In *Proceedings of the 2016 International Conference on Supercomputing*. 1–12.
- [76] Kiran Matam, Siva Rama Krishna Bharadwaj Indarapu, and Kishore Kothapalli. 2012. Sparse matrix-matrix multiplication on modern architectures. In *2012 19th International Conference on High Performance Computing*. IEEE, 1–10.
- [77] Ahmed A Abdelrahman, Mohamed M Fouad, Hisham Dahshan, and Ahmed M Mousa. 2017. High performance CUDA AES implementation: A quantitative performance analysis approach. In *2017 Computing conference*. IEEE, 1077–1085.
- [78] Gert-Jan van den Braak, Bart Mesman, and Henk Corporaal. 2010. Compile-time GPU memory access optimizations. In *2010 International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation*. IEEE, 200–207.
- [79] Michael Bauer, Henry Cook, and Brucec Khailany. 2011. CudaDMA: optimizing GPU memory bandwidth via warp specialization. In *Proceedings of 2011 international conference for high performance computing, networking, storage and analysis*. 1–11.
- [80] Takumi Honda, Yasuaki Ito, and Koji Nakano. 2015. A warp-synchronous implementation for multiple-length multiplication on the GPU. In *2015 Third International Symposium on Computing and Networking (CANDAR)*. IEEE, 96–102.
- [81] Joseph D Garvey and Tarek S Abdelrahman. 2018. A strategy for automatic performance tuning of stencil computations on GPUs. *Scientific Programming* 2018, 1 (2018), 6093054.
- [82] Da Yan, Wei Wang, and Xiaowen Chu. 2020. Demystifying tensor cores to optimize half-precision matrix multiply. In *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 634–643.
- [83] Md Mehrab Hossain Opi, Sumaiya Khan, and Moshhammad Farzana Rahman. 2025. Accelerating Bangla NLP Tasks with Automatic Mixed Precision: Resource-Efficient Training Preserving Model Efficacy. *arXiv preprint arXiv:2512.00829* (2025).
- [84] Enkhbold Nyamsuren. 2024. Evaluating Quantized Large Language Models for Code Generation on Low-Resource Language Benchmarks. *arXiv preprint arXiv:2410.14766* (2024).
- [85] Pawel K Radtke and Tobias Weinzierl. 2025. Compiler-supported reduced precision and AoS-SoA transformations for heterogeneous hardware. *arXiv preprint arXiv:2512.05516* (2025).
- [86] Junlong Li, Daya Guo, Dejian Yang, Runxin Xu, Yu Wu, and Junxian He. 2025. Codei/o: Condensing reasoning patterns via code input-output prediction. *arXiv preprint arXiv:2502.07316* (2025).
- [87] Yangruibo Ding, Jinjun Peng, Marcus Min, Gail Kaiser, Junfeng Yang, and Baishakhi Ray. 2024. Semcoder: Training code language models with comprehensive semantics reasoning. *Advances in Neural Information Processing Systems* 37 (2024), 60275–60308.
- [88] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515* (2024).
- [89] Mingzhe Du, Luu Anh Tuan, Yue Liu, Yuhao Qing, Dong Huang, Xinyi He, Qian Liu, Zejun Ma, and See-kiong Ng. 2025. Afterburner: Reinforcement learning facilitates self-improving code efficiency optimization. *arXiv preprint*

1373 *arXiv:2505.23387* (2025).
1374 [90] Jian Yang, Xianglong Liu, Weifeng Lv, Ken Deng, Shawn Guo, Lin Jing, Yizhi Li, Shark Liu, Xianzhen Luo, Yuyu Luo,
1375 et al. 2025. From Code Foundation Models to Agents and Applications: A Comprehensive Survey and Practical Guide
1376 to Code Intelligence. *arXiv preprint arXiv:2511.18538* (2025).
1377 [91] Yash Akhauri, Xingyou Song, Arissa Wongpanich, Bryan Lewandowski, and Mohamed S Abdelfattah. 2025. Regression
1378 language models for code. *arXiv preprint arXiv:2509.26476* (2025).
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421