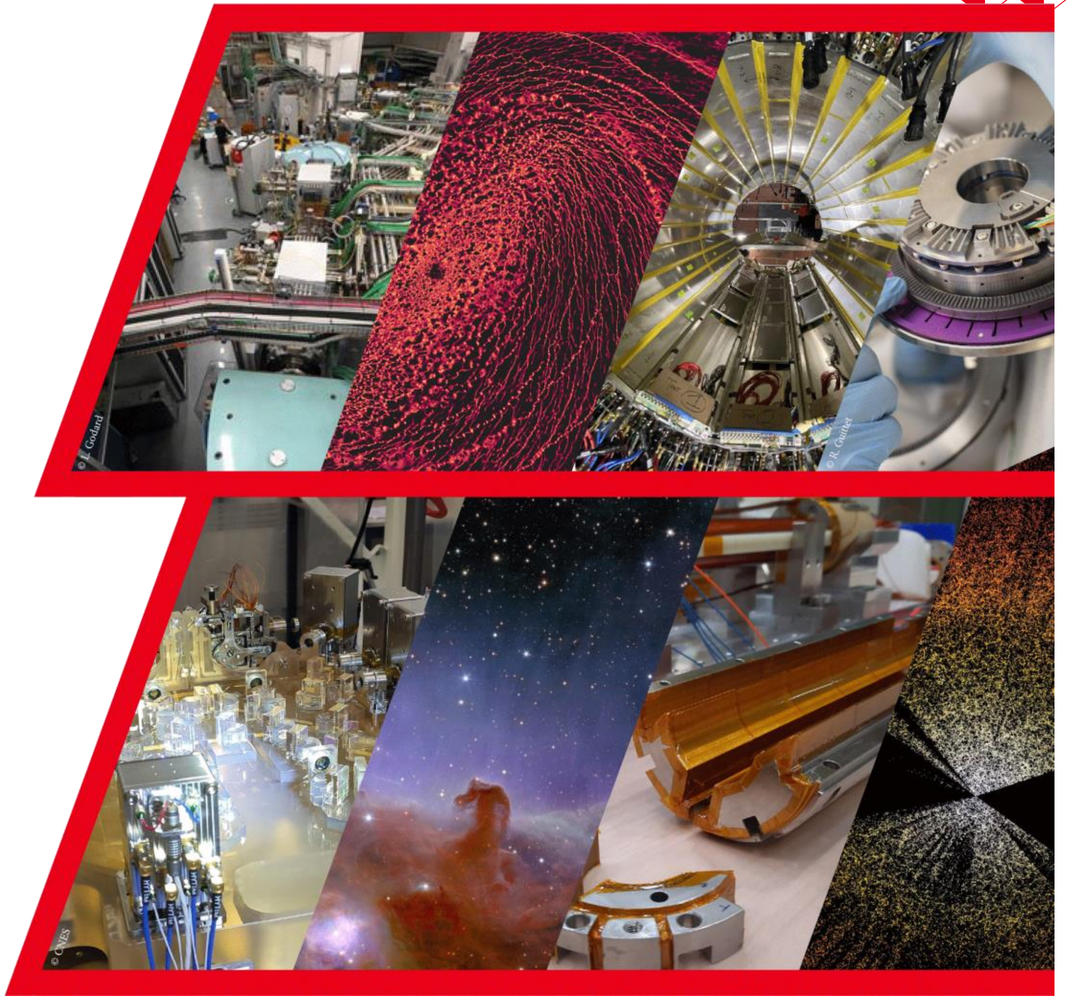
MM/LLMs in Scientific Research and High-Performance Computing



Imed MAGROUNE – 2025/02/19

Imed.magroune@cea.fr

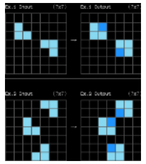
The French Alternative Energies and Atomic Energy Commission (CEA)

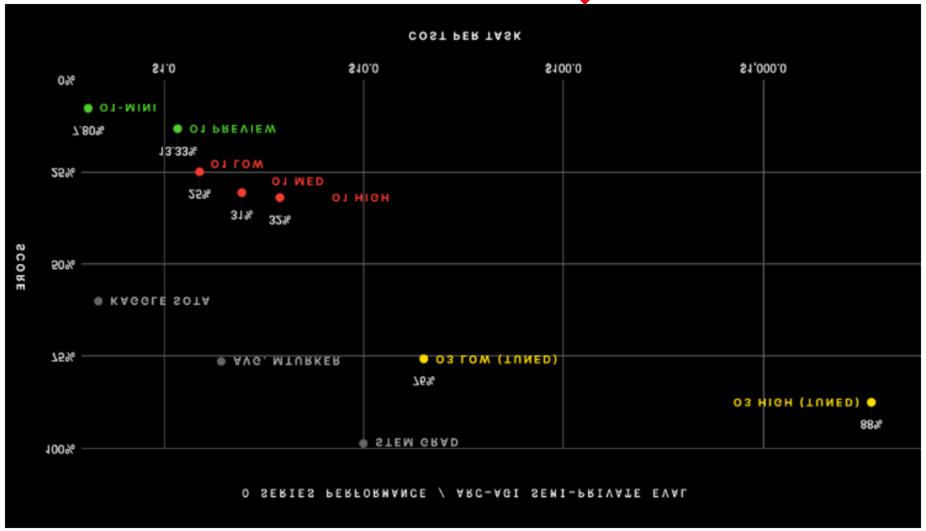
GenAI: Transforming Scientific Research

- AI Models in Research:
  - Classification and prediction
  - Computer Vision
  - Pattern recognition
  - Automation of specific tasks
  - .. etc

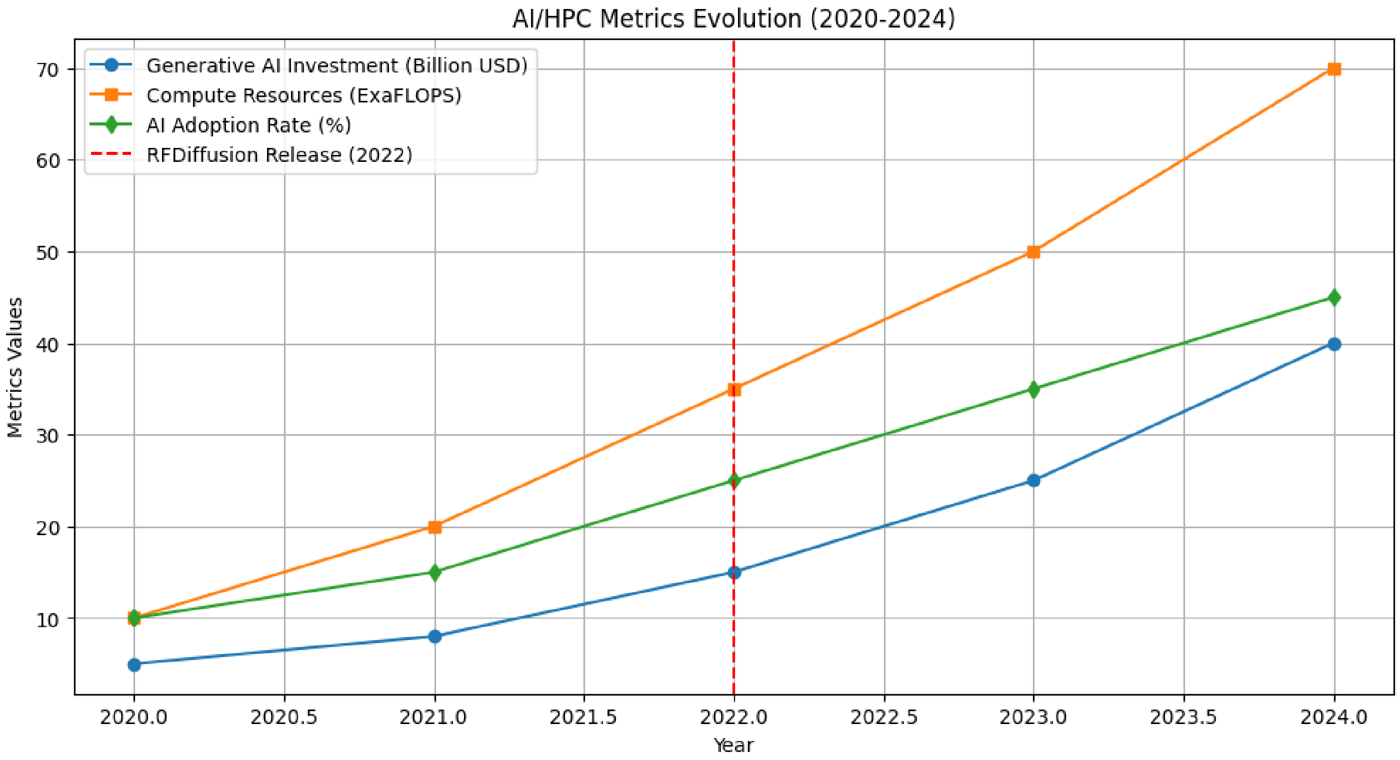
Generative AI represents a paradigm shift in scientific research, fundamentally changing how we approach complex computational challenges and scientific discovery.

- The Missing Piece: General Intelligence, from knowledge to reasoning Models
- Toward AGI –*already reached-*





https://arcprize.org/



Src : AI Index Report (2024)   State of AI Report Compute Index:

State-of-the-Art LLMs

Historical Perspective

CERN's Early AI Success (2010-2020)

- Pattern recognition in particle physics
- Real-time data analysis in LHC experiments
- Machine learning for event reconstruction

Current SOTA Models (2025-02)

OpenAI GPT-5

- Multimodal capabilities
- Advanced reasoning in scientific contexts
- Performance benchmarks in scientific tasks

Deepseek R1

- advanced reasoning
- Mixture of Experts (MoE) architecture
- reinforcement learning approach
- Open source & open wieghts
- distilled **1B** model outperforms GPT-4o and Claide 3,5 !

Others :

- Google Gemini
- Mistral
- ….

$30 Deep Seek Breakthrough by UC Berkeley PhD Student

What is the "Aha Moment"?
The "aha moment" refers to a sudden realization or a breakthrough in problem-solving observed during the training of Deep Seek models. During this phase, the model demonstrates the ability to re-evaluate its initial approach and allocate additional time to complex problems.

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954

From the table, DeepSeek-R1-Distill-Qwen-1.5B outperforms GPT-4o and Claude-3.5 in specific tasks like:

From Traditional ML to State-of-the-Art LLMs

What can LLMs/MMLLMs do ?:

- Understanding context
- Reasoning capabilities
- Problem-solving abilities

Complementing Existing Models

- Enhanced data interpretation
- Intelligent preprocessing
- Results analysis and explanation

Integration Benefits:

- Combining specialized AI with general intelligence
- Enhanced decision-making
- Automated insight generation

Benchmark Comparisons:

- Scientific reasoning (MMLU scores)
- Code generation accuracy
- Mathematical problem-solving capability

Impact Metrics:

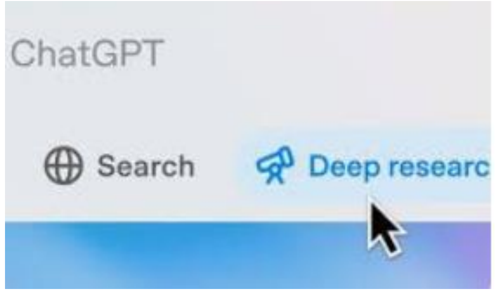
- 90%+ accuracy in scientific paper analysis
- 75% reduction in research time for literature review
- 60% improvement in code optimization tasks

A practical sample : *Deep Research takes question answering to the next level by spending 2-4 minutes doing the work it would take a human expert many hours to perform. Here's how it works:*

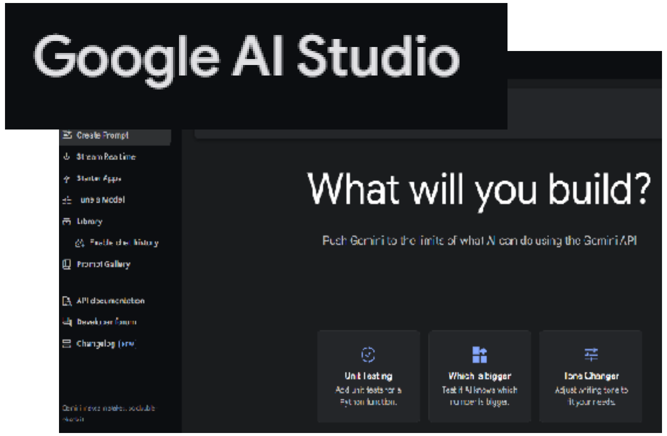
https://www.perplexity.ai/

- Research with reasoning and coding capabilities
- iteratively searches, reads documents, and reasons about what to do next
- refining its research plan as it learns more about the subject areas.
- Report writing, the agent then synthesizes all the research into a clear and comprehensive report.
- Export & Share pf final report to a PDF or document and share it with colleagues or friends.

Gemini

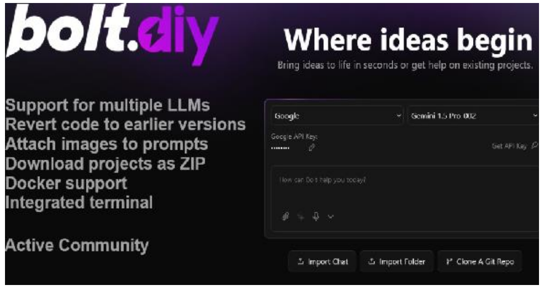


From Foundation models to smallers LLMs



220 $!
1028 Cuda core

Jetson Orin Nano Developer Kit NVIDIA



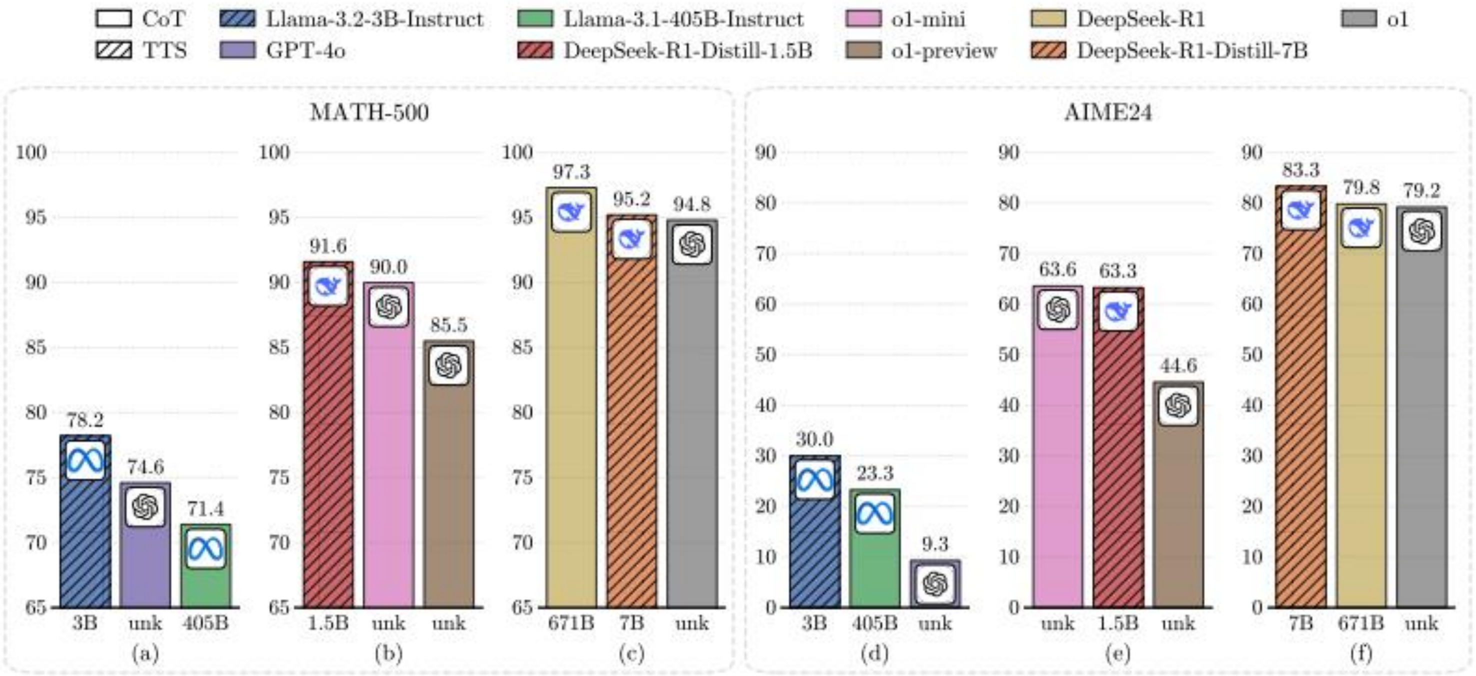


Figure 1: Comparison between the performance of smaller LLMs compute-optimal TTS and that of larger LLMs CoT on MATH-500 and AIME24. **(a) & (d)** Llama-3.2-3B-Instruct surpasses Llama-3.1-405B-Instruct and GPT-4o on MATH-500 and AIME24; **(b) & (e)** DeepSeek-R1-Distill-1.5B outperforms o1-preview on MATH-500 and AIME24, and surpasses o1-mini on MATH-500; **(c) & (f)** DeepSeek-R1-Distill-7B beats o1 on MATH-500 and AIME24, and exceeds DeepSeek-R1 on AIME24.

* Work done during an internship at Shanghai AI Laboratory
† Corresponding authors: Biqing Qi (qibiqing@pjlab.org.cn), Bowen Zhou (zhoubowen@tsinghua.edu.cn)

Can 1B LLM Surpass 405B LLM? Rethinking Compute-Optimal Test-Time Scaling :https://arxiv.org/pdf/2502.06703

Transforming Scientific Collaboration and Analysis

High-Precision Scientific Translation

- Cross-language accuracy exceeding 98% for technical content
- Real-time translation of research papers and technical documentation
- Domain-specific terminology management across 95+ languages
- Preservation of mathematical formulas and scientific notation
- Example: Claude 3's ability to translate and explain quantum physics papers across languages while maintaining technical accuracy

Automated Publication Analysis

- Processing 100,000+ papers per day across scientific databases
- Real-time trend analysis and research gap identification
- Citation network mapping and influence tracking
- Automated meta-analysis of research findings
- Example: Elsevier's AI tools analyzing COVID-19 research, processing 200,000+ papers in months instead of years

Enhanced Report Generation

- Automated synthesis of multi-source research findings
- Generation of publication-ready figures and tables
- Consistency checking across large datasets
- Real-time literature updates and incorporation
- Example: Nature's use of AI for preliminary paper screening, reducing review time by 60%

International Collaboration Enhancement

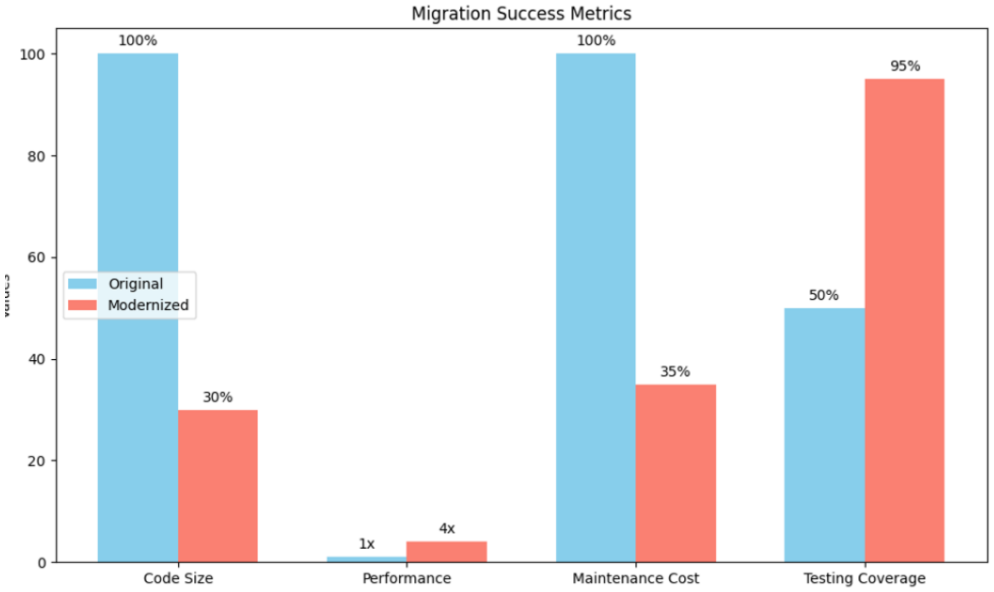
- Real-time multi-language research meetings
- Automated meeting summaries and action items
- Cross-cultural communication optimization
- Shared knowledge base development
- Example: CERN's AI-powered collaboration platform connecting 10,000+ scientists across 100 countries

70% reduction in literature review time
85% accuracy in technical translation
50% faster research paper drafting
3x increase in international collaboration efficiency

HPC Optimization

Legacy Code Modernization :

AI-Driven Scientific Code Transformation Success Stories



Climate Model Modernization

- NCAR's Community Earth System Model (CESM)
- 1M+ lines of legacy FORTRAN → Modern C++/Python
- 3x performance improvement
- Risk-free migration validated against 30 years of data

Nuclear Simulation Codes

- Lawrence Livermore Lab's Legacy Code Migration
- 40-year-old simulation codes successfully modernized
- 65% reduction in maintenance costs
- Zero loss of precision in critical calculations

Key Projects and Tools

Microsoft's AI-Powered Code Migration Suite
Google's FORTRAN-to-Python Translator
OpenAI Codex for Scientific Computing
Anthropic's Claude for Code Analysis

HPC Optimization

Modernizing Scientific Computing
Legacy Code Migration

- Automated analysis of legacy FORTRAN/COBOL code
- AI-driven code modernization and optimization
- Risk-free transformation of critical applications
- Example: Successful migration of 30-year-old climate models

Performance Enhancement

- AI-driven code parallelization
- Automatic GPU optimization
- Memory usage optimization
- Example: 40% performance gain in molecular dynamics simulations

Architecture Adaptation

- Code adaptation for modern HPC architectures
- Automatic scaling for cloud environments
- Energy efficiency optimization
- Example: Auto-tuning for exascale computing

From Supercomputers to Edge Devices
Low-Level Optimizations

- SIMD Acceleration
  - AVX-512 optimization for scientific kernels
  - Automatic vectorization with AI assistance
  - Performance gains:
  - 4-8x speedup in linear algebra operations
  - 3x in FFT computations
  - 6x in molecular dynamics kernels

- Rust's Rayon for Scientific Computing
  - Data-parallel computations
  - Zero-cost abstractions
  - Success stories:
  - 5x speedup in genomics analysis
  - 3x faster protein folding calculations
  - Automatic thread scaling

- CUDA Optimization
  - AI-driven kernel optimization
  - Automatic memory management
  - Recent developments

- CUDA Graph optimization
  - Multi-GPU scaling
  - Dynamic kernel fusion

No-Code and AI-Driven Development

Transforming Scientific Programming
AI Agents for Scientific Coding

GitHub Copilot Enterprise for Scientific Computing

- Specialized in mathematical algorithm implementation
- Understanding of complex scientific notation
- Integration with scientific libraries (NumPy, SciPy, etc.)

Code Generation Evolution

- Automated translation of scientific formulas to code
- Context-aware code suggestions based on research papers
- Multi-language support (FORTRAN → Python/C++ conversion)
- **Example: Claude 3 converting mathematical papers directly to executable code**

Documentation and Knowledge Preservation

- Auto-documentation of legacy scientific code
- Preservation of domain expertise in AI models
- Knowledge extraction from retiring developers
- Example: DeepMind's AlphaCode documenting complex algorithms

NoOps Revolution

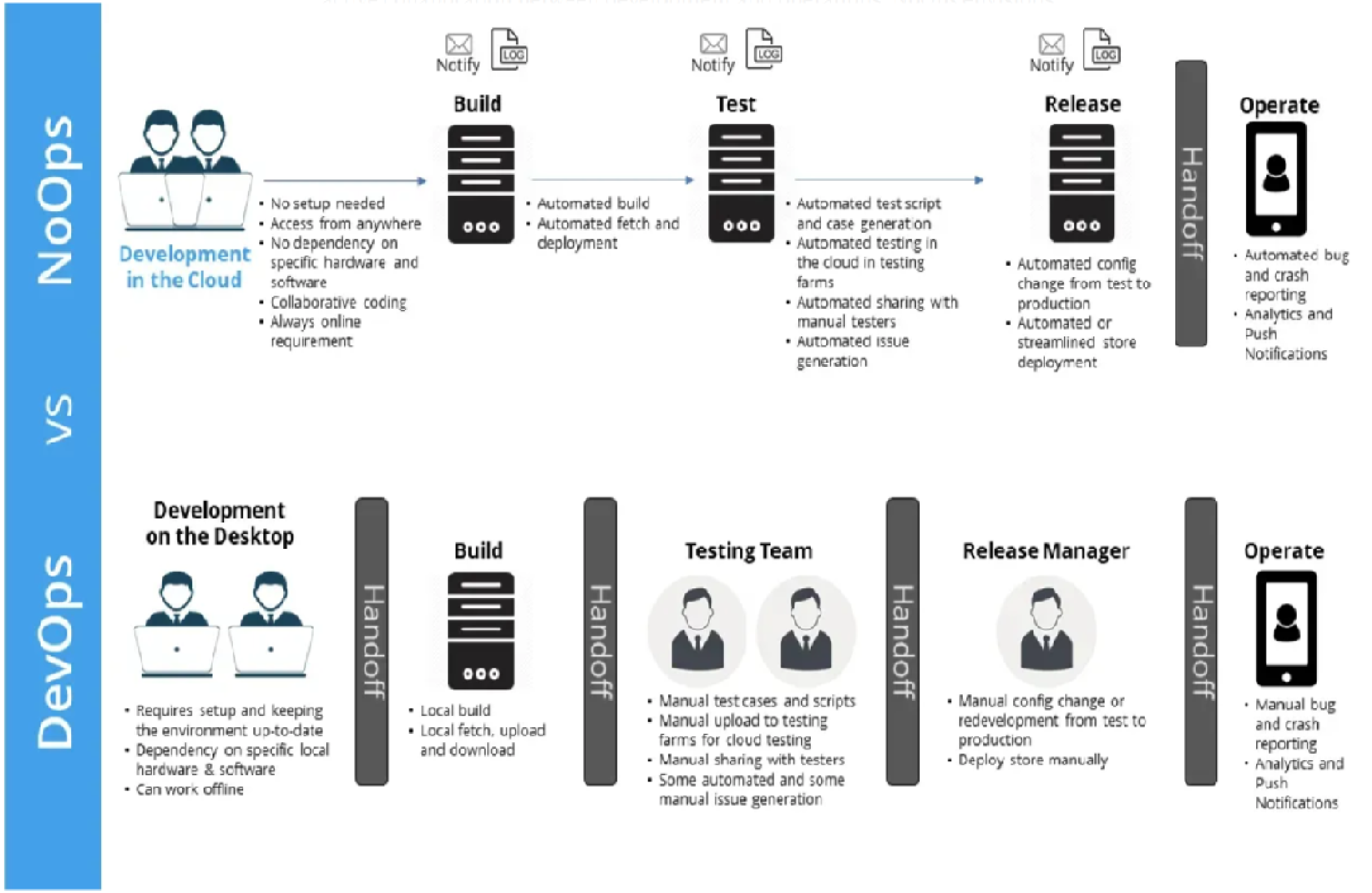
Beyond Traditional DevOps

AI-Powered Infrastructure

- Self-healing systems for HPC environments
- Automated resource allocation and scaling
- Real-time performance optimization
- Example: Google's Cloud AI optimizing research workloads

Intelligent Operations

- Predictive maintenance of computing clusters
- Automated security patching and compliance
- Energy consumption optimization
- AI agents managing entire compute environments



Future Perspectives

The New Era of Scientific Computing
Transformative Technologies

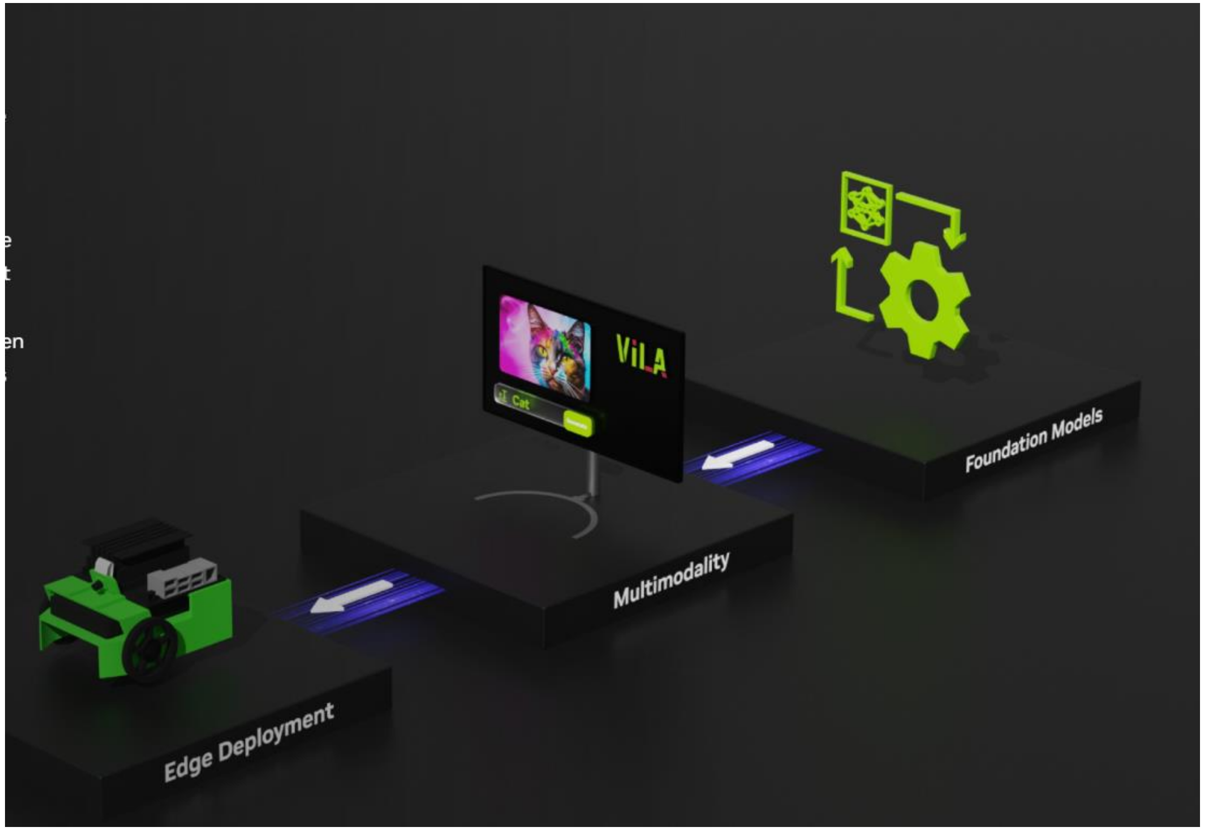
AI Research Assistants
- Autonomous experiment design
- Automated literature review
- Hypothesis generation and testing

Skills Evolution
From coding to AI prompt engineering
Focus on scientific thinking over implementation
Hybrid AI-Human research teams

Research Acceleration
- 10x faster research cycles
- Democratized access to advanced computing
- Global collaboration networks
- Example: Open Science initiatives powered by AI

Key Trends to Watch
- Quantum-AI integration
- Automated scientific discovery
- AI-driven peer review
- Real-time research collaboration

Democratized HPC
- Turnkey solutions for small labs
- Automated optimization
- Cloud-edge hybrid models



70% reduction in legacy code migration time
40-60% improvement in code performance
80% reduction in operational incidents
5x acceleration in research cycles