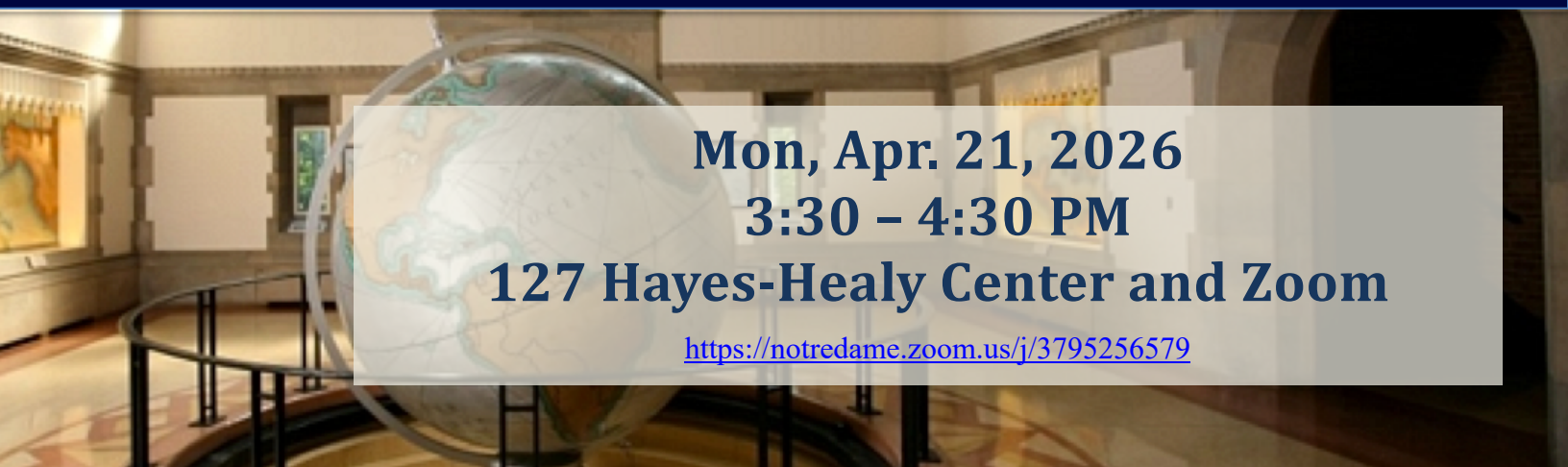
DAC WorkGroup on Scientific AI Agents
University of Notre Dame

Vansh Sharma

University of Michigan

Beyond Plausible Code: First-Principles Agentic Optimization for Scientific Computing

Large language model (LLM) agentic systems show promise for scientific code generation and design optimization, but their reliability remains uncertain in high-stakes domains that require strict physical, mathematical, and performance constraints. We present two complementary frameworks to address this challenge: Chain of Unit-Physics, a first-principles, multi-agent approach that encodes expert knowledge as unit-physics tests to constrain code generation, and AUTO, an iterative optimization framework that treats design as a gradient-free strategic search guided by LLM reasoning. We evaluate these methods on scientific and GPU computing benchmarks, including a 12-degree-of-freedom combustion solver, chemical kinetics kernels, matrix multiplication, and KernelBench tasks. Standard closed-weight and code-focused agents frequently fail to produce correct end-to-end solutions, exhibiting interface hallucinations, overconfident assumptions, numerical or physical inconsistency, configuration fragility, and occasional performance cheating. In contrast, our frameworks improve both robustness and performance: Chain of Unit-Physics converges in 5–6 iterations, matches a human-expert combustion implementation with a mean error of 3.1e−3 %, and achieves about 33.4% faster runtime and 30% lower memory use; AUTO delivers up to 1.74× speedup over in-lab optimized chemical kinetics code, reaches 94% of cuBLAS double-precision performance for matrix multiplication, and attains up to 118× speedups over PyTorch baselines on KernelBench. A posteriori analysis further shows that AUTO's search behavior aligns 50-70% with Bayesian optimization sampling strategies. Together, these results show that while zero-shot LLM code generation alone remains unreliable for scientific applications, embedding first-principles validation and iterative multi-agent search yields a practical, interpretable, and cost-effective path toward trustworthy human–AI collaboration in scientific computing.



Mon, Apr. 21, 2026
3:30 – 4:30 PM
127 Hayes-Healy Center and Zoom

https://notredame.zoom.us/j/3795256579