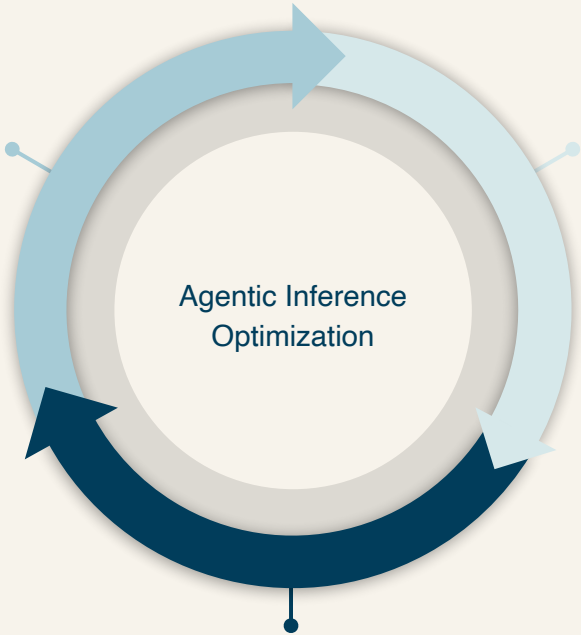
Agentic Solutions For Inference Optimization

2026 MLSys Young Professionals Symposium

Kel Vanee, MVP, Machine Learning Engineering

Agentic inference optimization



Set Inference Objectives

Define Model, Infrastructure and Performance Thresholds

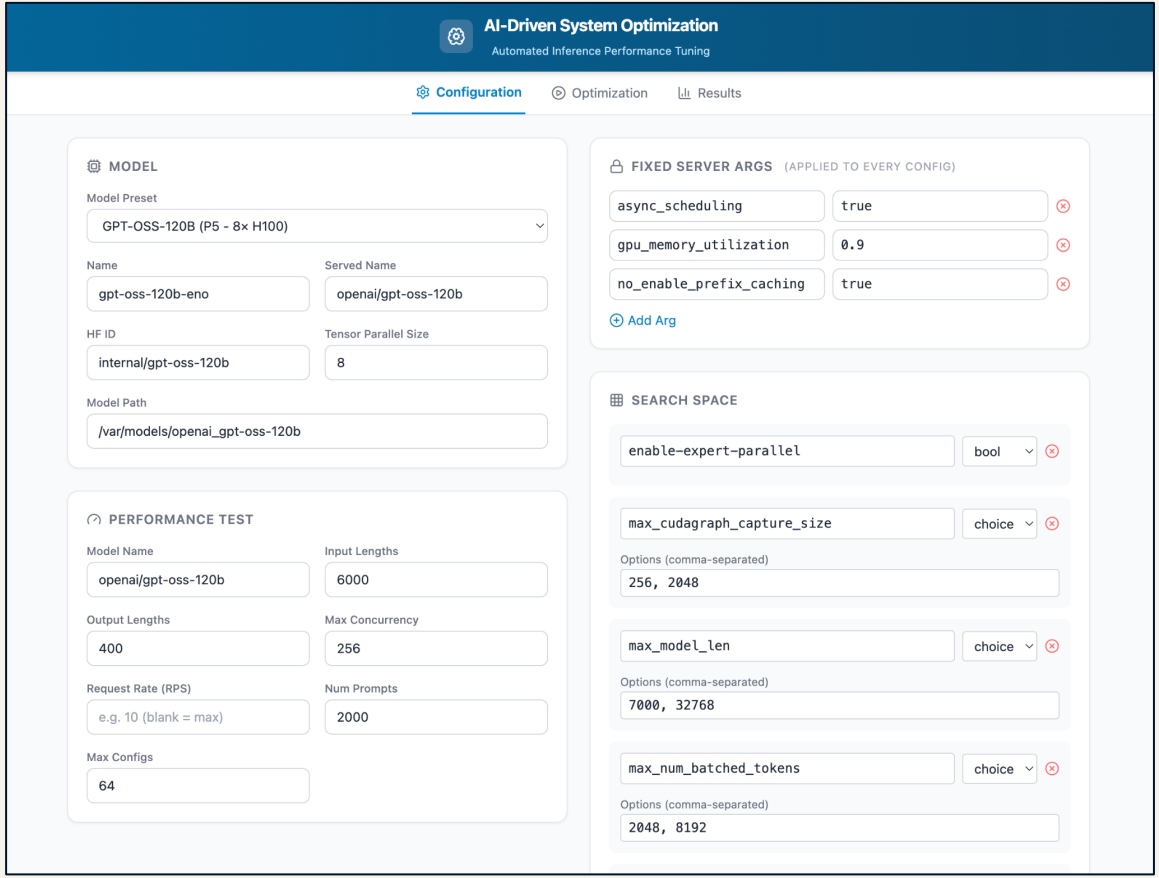
Auto-deploy configurations and Benchmark

Create deployments from a configuration search space and collect performance metrics

Recommend Optimal Configurations

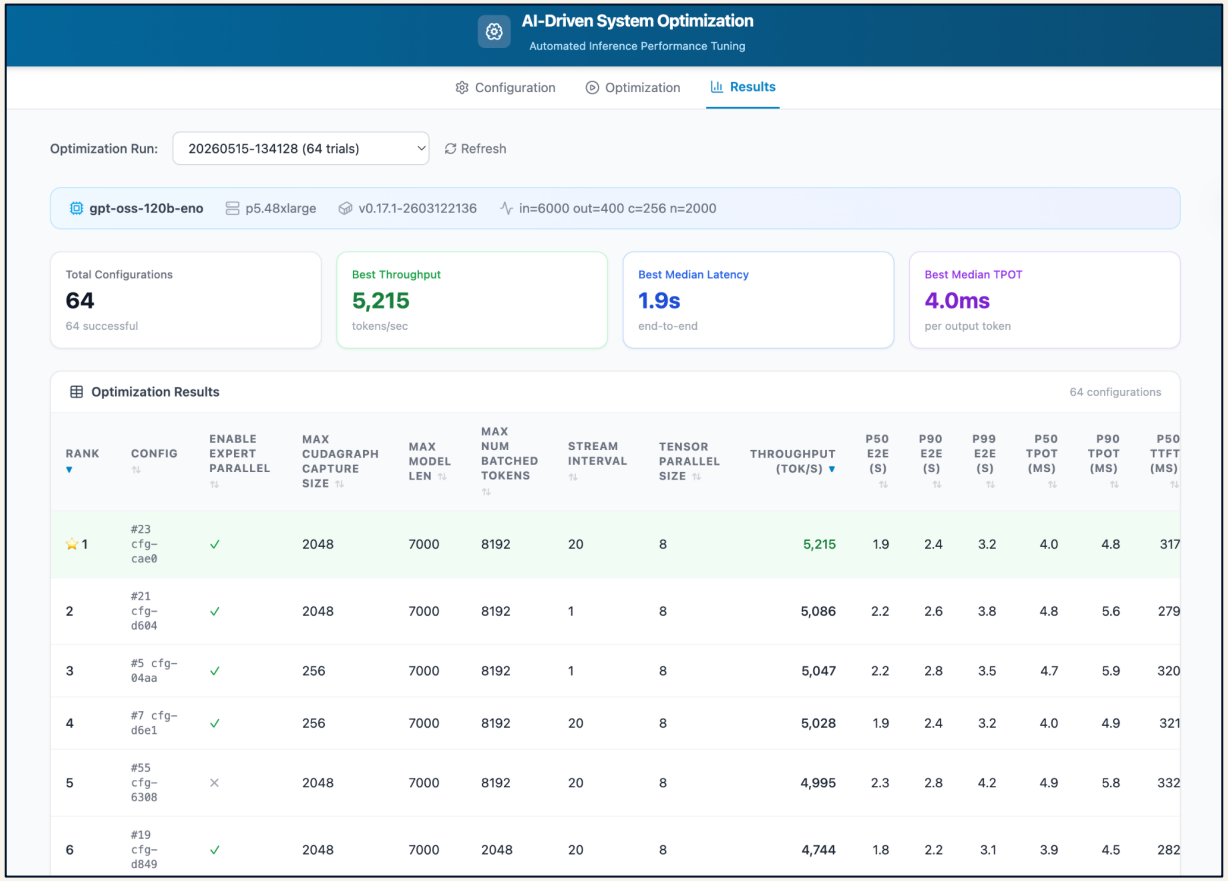
Analyze benchmark results to predict optimal settings

Scale Engine Tuning Through Automation



Note: the above represents sample data and is provided for illustrative purposes only

Scale Engine Tuning Through Automation



RANK	CONFIG	ENABLE EXPERT PARALLEL	MAX CUDAGRAPH CAPTURE SIZE	MAX MODEL LEN	MAX NUM BATCHED TOKENS	STREAM INTERVAL	TENSOR PARALLEL SIZE	THROUGHPUT (TOK/S)	P50 E2E (S)	P90 E2E (S)	P99 E2E (S)	P50 TPOT (MS)	P90 TPOT (MS)	P50 TTFT (MS)
☆ 1	#23 cfg-cae0	✓	2048	7000	8192	20	8	5,215	1.9	2.4	3.2	4.0	4.8	317
2	#21 cfg-d604	✓	2048	7000	8192	1	8	5,086	2.2	2.6	3.8	4.8	5.6	279
3	#5 cfg-04aa	✓	256	7000	8192	1	8	5,047	2.2	2.8	3.5	4.7	5.9	320
4	#7 cfg-d6e1	✓	256	7000	8192	20	8	5,028	1.9	2.4	3.2	4.0	4.9	321
5	#55 cfg-6308	✗	2048	7000	8192	20	8	4,995	2.3	2.8	4.2	4.9	5.8	332
6	#19 cfg-d849	✓	2048	7000	2048	20	8	4,744	1.8	2.2	3.1	3.9	4.5	282

Note: the above represents sample data and is provided for illustrative purposes only

Capital One AI Research **Internships**

We can’t do it without great talent like you. Join us.

WANT TO LEARN MORE?
MEET US AT BOOTH #8

Applied Research Internship Program (ARIP)

Ph.D. students can expand their AI knowledge and move their research forward through Capital One’s Applied Research Internship Program. During the 12-week summer internship, you’ll engage in high-impact applied research, exploring novel and interesting AI challenges that will help create transformative customer experiences.

Data Science Internship Program (DSIP)

Tap into the latest computing and machine learning tech in this program for Master’s and PhD students. During the 10-week summer internship, you’ll create dynamic models and leverage a diverse tech stack as you work across teams to unlock opportunities that improve the lives of our customers.

AI Engineering Internship Program

Master’s and PhD students can be part of a team delivering industry-leading capabilities and scalable, high-performance AI infrastructure. During this 12-week program, you'll help build AI and Agentic AI applications and platform capabilities that power customer facing experiences, cutting-edge GenAI workflows, shared infrastructure, and internal developer tools.