

Scaling On-Premises Agentic Workloads from Deskside to Datacenter with the Dell AI Factory with NVIDIA

KEY RESULTS

DESKSIDE SOLUTIONS

UP TO

90%

vs AWS Bedrock

UP TO

98%

vs leading frontier API

DATACENTER SOLUTIONS

UP TO

68%

vs AWS Bedrock

UP TO

94%

vs leading frontier API

8-98%

savings range across every configuration evaluated

~2 months

fastest payback vs AWS Bedrock

30,000+

concurrent agents on the largest configuration evaluated

WORKLOAD SAVINGS

Evaluated across 9 Dell AI Factory platforms · 27 configurations · models from 30B to 120B parameters

Knowledge Worker
Big Savings from Lightweight AI

NVIDIA Nemotron™ 3 Nano · 3.6B/31.6B

26 → 33,821 agents

vs Bedrock

8-90%

vs Leading Frontier API

86-98%

Up to \$9.2M saved vs AWS Bedrock · Up to \$109M saved vs leading frontier APIs

Sales Agent
Balanced Complexity, Consistent Savings

50/50 NVIDIA Nemotron™ 3 Nano + NVIDIA Nemotron™ 3 Super

97 → 8,676 agents

vs Bedrock

41-85%

vs Leading Frontier API

89-97%

Up to \$5.4M saved vs AWS Bedrock · Up to \$50.6M saved vs leading frontier APIs

Software Development
Maximum Complexity, Maximum Savings

NVIDIA Nemotron™ 3 Super · 12B/120B

2 → 4,850 agents

vs Bedrock

8-88%

vs Leading Frontier API

82-98%

Up to \$9.7M saved vs AWS Bedrock · Up to \$68.8M saved vs leading frontier APIs

TIME TO BREAK EVEN, BY PLATFORM (KEY CONFIGURATIONS)

