From Benchmark Gaps to Domain-Aware AI Systems

EPRI Open Power AI Domain Specific Model / Domain Intelligence Working Group

(joint session with the OPAI Implementation Working Group)

Apurba Sakti, *Ph.D.*
Electric Power Research Institute
asakti@epri.com

Monthly Working Group Meeting

18 Jun 2026

TOKEN HYGIENE

Use AI with intent.

Share the right data, in the right tool, with the right controls.

AI can accelerate insight, productivity, and collaboration—if we handle data deliberately.

AI SAFETY SHARE

WHY THIS MATTERS

50%
of U.S. employees use AI in their role

66%
of remote-capable employees use AI at work

42%
report non-public company information has been entered into GenAI tools

63%
of organizations lack AI governance policies to manage AI deployment or prevent unauthorized AI use

PAUSE AND ASK BEFORE YOU PASTE OR UPLOAD

IS IT APPROPRIATE TO SHARE?

Is the information public, or does it include confidential, proprietary, export-controlled, personal, employee, or otherwise restricted information?

IS THIS THE RIGHT TOOL?

Do you know where the data is going, how it is stored, whether it is retained, and whether it can be used to train models?

When in doubt, use an approved environment or leave it out.

VERIFY HIGH-IMPACT OUTPUTS

AI can be helpful and still be wrong. Verify important facts, numbers, sources, and technical conclusions.

MANAGE MEETING DATA

Use recordings, transcripts, AI notetakers, and summaries only when appropriate and with awareness of privacy, consent, and data-sharing rules.

GOOD TOKEN HYGIENE ENABLES TRUSTED AI USE

The goal is not to avoid AI. It is to use AI responsibly so we can benefit from it safely, securely, and credibly.

PUBLIC SESSION REMINDER

Please keep the discussion and any shared materials suitable for public release, as materials may be shared through the Open Power AI website.

openpowerai.org

THINK BEFORE YOU PASTE.
USE AI RESPONSIBLY. PROTECT WHAT MATTERS.

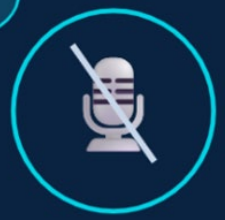
Be intentional | Be responsible | Be respectful | Be secure

Sources: Gallup, Rising AI Adoption Spurs Workforce Changes (2026) and *Frequent Use of AI in the Workplace Continued to Rise in Q4* (2026); **Cisco**, 2025 Data Privacy Benchmark Study; **IBM**, Cost of a Data Breach Report 2025.

Housekeeping

A few norms to keep today’s discussion useful, interactive, and publicly shareable

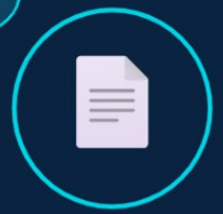
1



Participate actively

Please mute upon entry. Use raise hand or chat for questions and discussion.

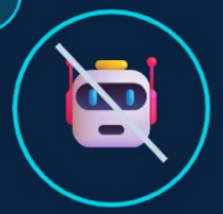
2



Keep it publicly shareable

Please share only material suitable for public release.

3



No automated notetakers

Please do not add automated notetakers to this meeting.

4

Recording notice

This session will be recorded. Participation acknowledges consent.

5

Goal for today

Practical discussion and shared learning. Product pitches should be avoided or kept at a minimum.

Powered by EPRI

OPEN POWER AI CONSORTIUM

Our Work Groups

Domain-specific model

Domain-specific model trained on energy-specific datasets

Use Case Sandbox

Identify, prioritize, and evaluate real-world applications

Implementation

Share methods & best practices for AI deployment and scaling

300+ Global Members

Harnessing AI's transformative potential for a resilient, cost-effective energy future through collaboration

Open-source data and AI models for industry-wide value

NVIDIA. Google databricks

Microsoft ORACLE aws

Learn More:
www.openpowerai.org

Domain-Specific Model
to include knowledge from
> 15,000
EPRI Reports

250+
use cases identified for consortium prioritization and development

Join us in Shaping the Future of AI and Energy

Today's throughline: from benchmarking gaps to domain-aware AI systems

How benchmarking, knowledge augmentation, and model customization may connect to reliable AI performance in power-sector applications

1 Benchmarking reveals the gaps

Framing the session

- Identifies where general-purpose models perform well – and where they struggle
- Provides repeatable evidence to guide improvements

2 Domain Knowledge Grounding

Chris Trueblood
Principal Technical Leader, EPRI

Connect AI to organizational data, documents, tools, and workflows.

3 Adapt models for specialized domains

Jihyun Yang, *Ph.D.*
Solutions Architect, NVIDIA

Improving AI performance for specialized technical domains

4 Enable domain intelligence

Working Group Discussion

- Combine model capability + grounding + evaluation + controls
- Move towards more reliable, trusted AI for power-sector use cases

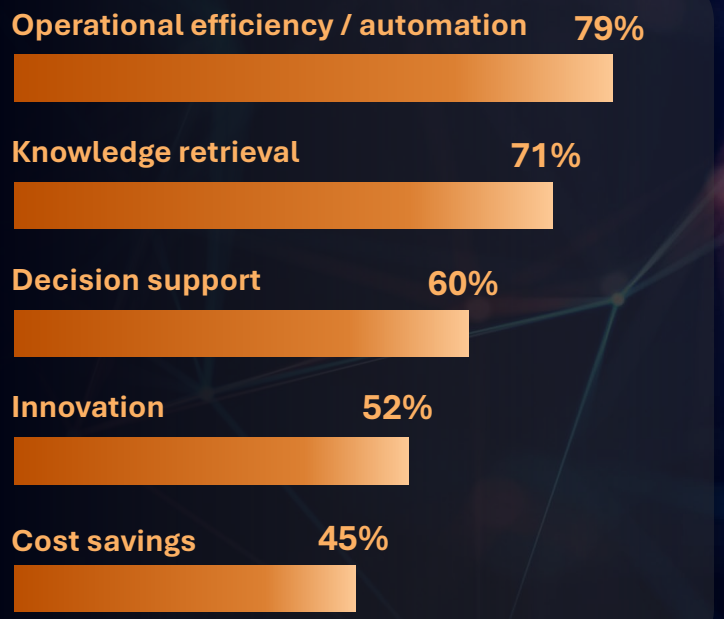
Takeaway

Better benchmarking identifies where gaps exist; domain-aware system design and model adaptation create the path to close them.

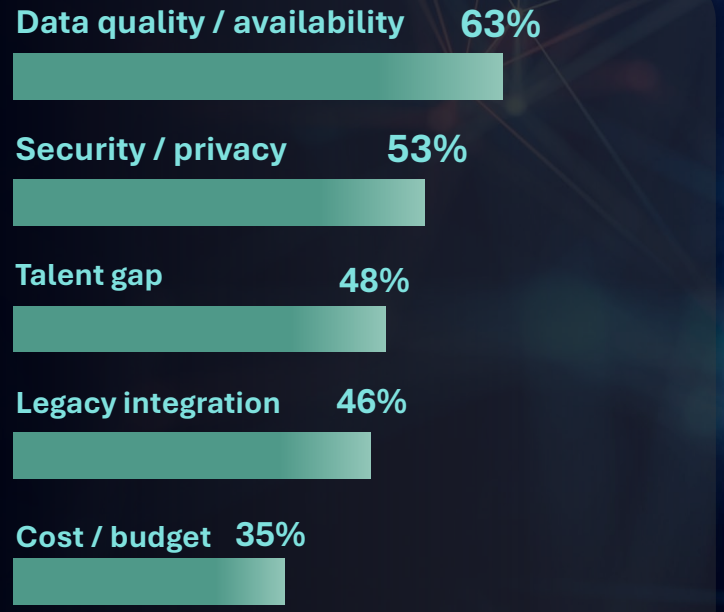
The electric power ecosystem sees value from AI; scaling needs include overcoming data, security, and integration challenges

OPAI Ongoing *Member Representative Committee Survey* Results

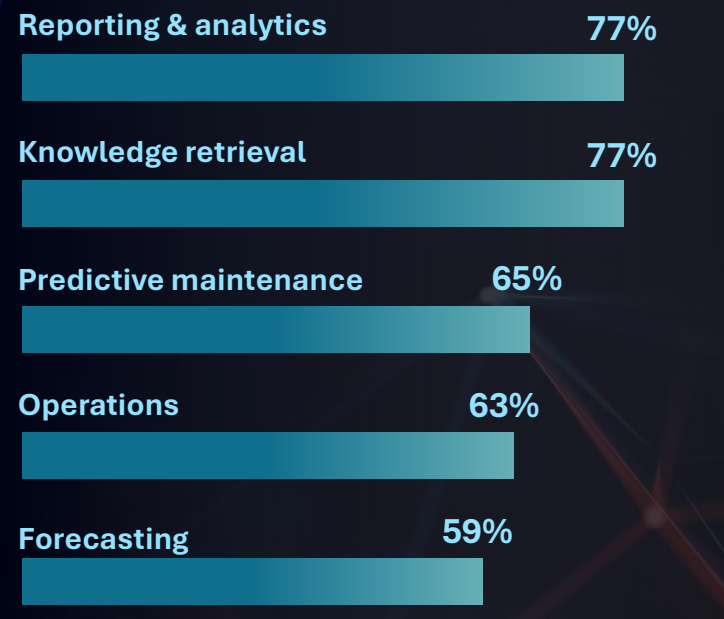
Where AI is delivering value



Top barriers to scaling



Where expansion likely in 1-3 years



Number of respondents vary from 131-133 with strong representation from generation, transmission, distribution, and AI solution providers

Implication To scale responsibly, **assurance evidence** (performance + controls + monitoring) is needed

AI adoption requires safeguards to scale responsibly & safely

External assurance frameworks converge on documented, verifiable assurance evidence

External frameworks emphasize assurance evidence

NIST AI RMF
Govern-Map-Measure-Manage
Measure focuses on TEVV + monitoring.

Other relevant frameworks

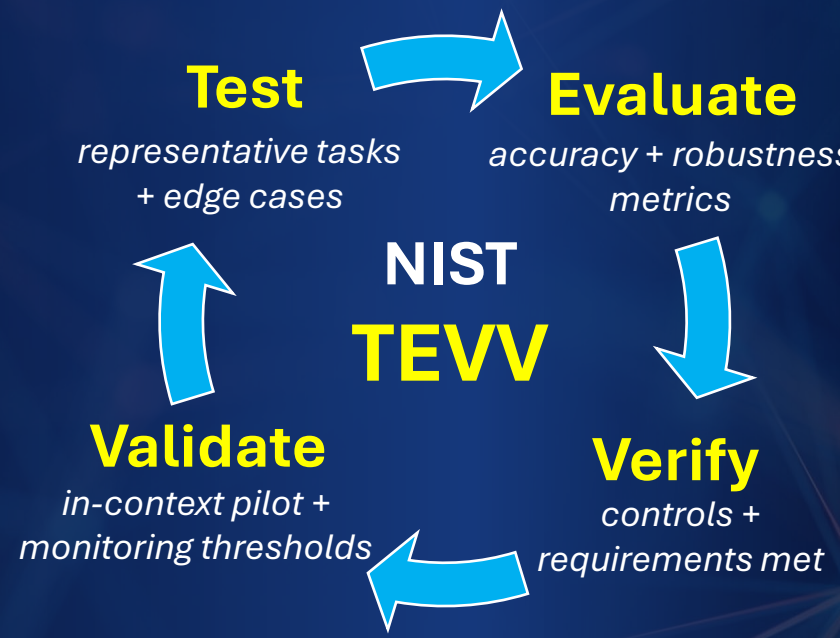
EU AI Act
Demonstrate robustness/accuracy + human oversight

ISO/IEC 42001 & ISO/IEC 23894
Monitoring, measurement, and auditability

IEC 62443 & NERC CIP
Security assurance: verified controls + monitoring

Evidence Loop

One practical way to organize assurance evidence



How EPRI's benchmarking aligns

Comparable: same tests/metrics across models/systems

Repeatable: versioned test suites + multi-run results

Actionable: thresholds + gaps that guide deployment choices

Assurance evidence = demonstrated performance + verified controls + validated operation

Takeaway

Frameworks set expectations; benchmarking provides repeatable, comparable evidence that helps move AI from pilots toward production

WattWorks.EPRI.Com

The Power Sector's AI Benchmarking Hub

The Power Sector's AI Benchmarking Hub

How Today's Leading LLMs Perform on Power-Sector Questions

EPRI benchmarks leading large language models on **real-world, utility-relevant questions** to help the power sector understand where today's AI systems are strong, **where they fall short**, and how performance evolves over time.

Live Benchmark — Updated as new models are evaluated

Download White Paper → Watch Benchmarking Discussion →

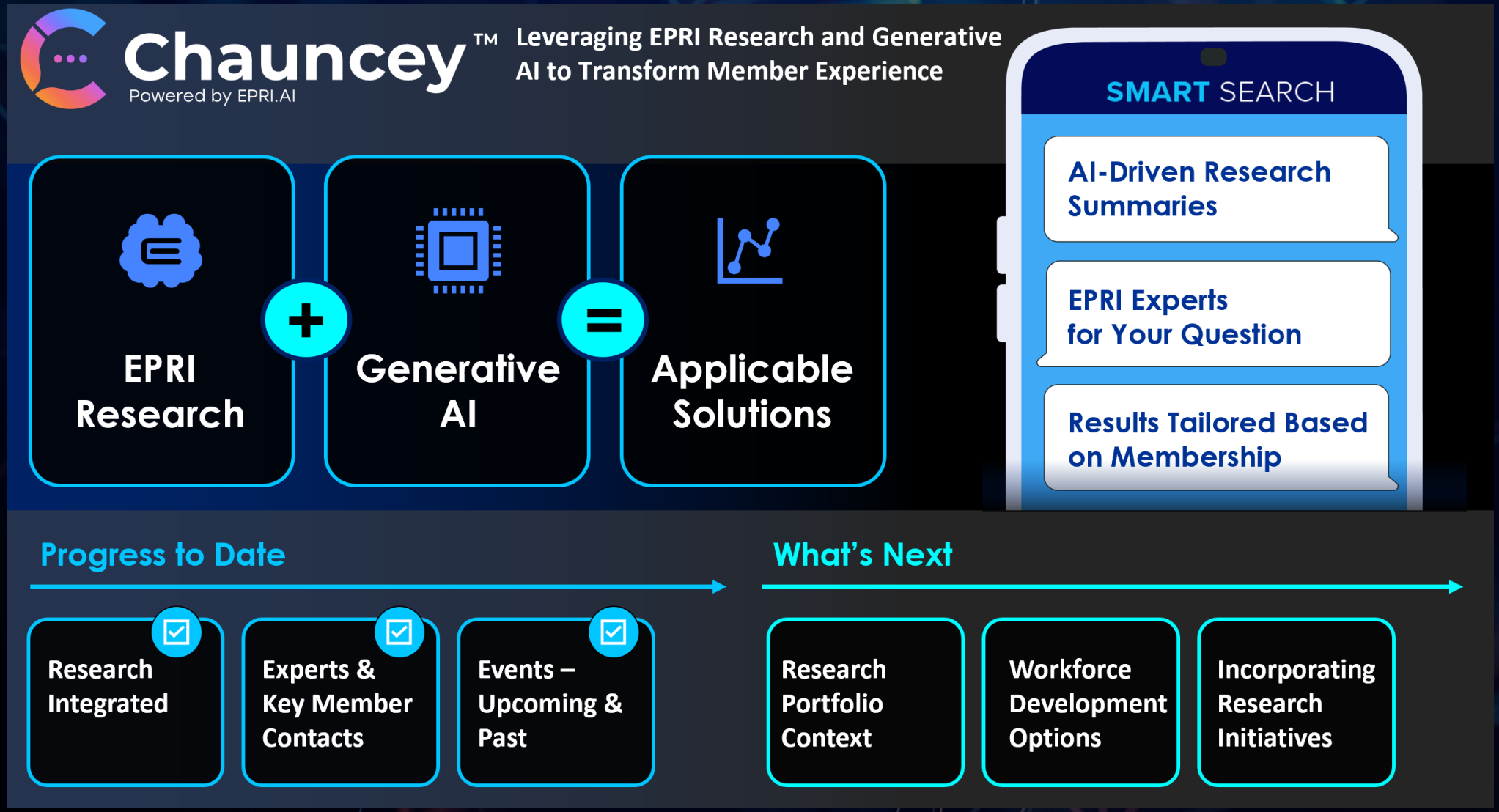
Multiple-choice evaluations | Open-ended short answers | Multi-run repeatability | Domain-augmented tools

INTERACTIVE BENCHMARK

Explore how selected models perform across EPRI benchmark datasets, including **short-answer** and **multiple-choice questions, model-only and web-search modes**, and breakdowns by **difficulty or topic**.

Use-case benchmarking in action: Knowledge retrieval

Chauncey™ as an example workflow: measuring grounded answers & citation quality for **research-backed insights**



Chauncey QR Code

Broader access to EPRI through the Microsoft ecosystem

Extending reach beyond 1:1 support

Domain Intelligence Working group

Future Discussion Topics – Please sign up

https://forms.office.com/r/AfcAdPDWgX

Today's throughline: from benchmarking gaps to domain-aware AI systems

How benchmarking, knowledge augmentation, and model customization may connect to reliable AI performance in power-sector applications

1 Benchmarking reveals the gaps

Framing the session

- Identifies where general-purpose models perform well – and where they struggle
- Provides repeatable evidence to guide improvements

2 Domain Knowledge Grounding

Chris Trueblood
Principal Technical Leader, EPRI

Connect AI to organizational data, documents, tools, and workflows.

3 Adapt models for specialized domains

Jihyun Yang, *Ph.D.*
Solutions Architect, NVIDIA

Improving AI performance for specialized technical domains

4 Enable domain intelligence

Working Group Discussion

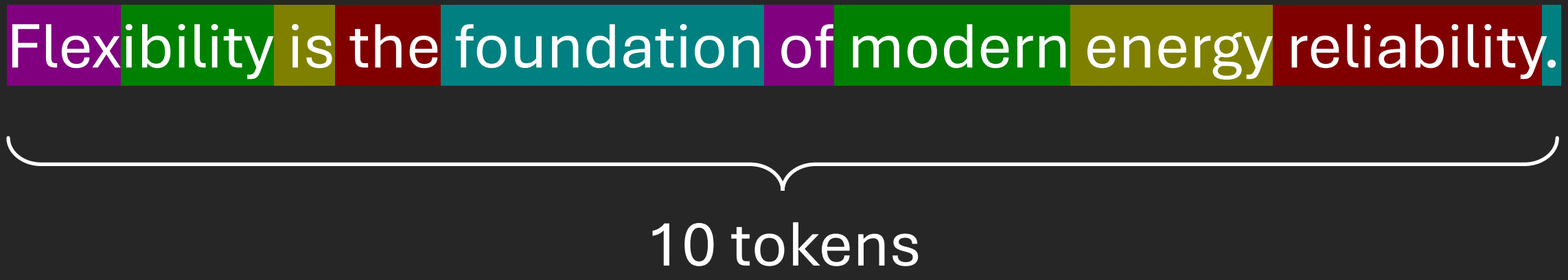
- Combine model capability + grounding + evaluation + controls
- Move towards more reliable, trusted AI for power-sector use cases

Takeaway Better benchmarking identifies where gaps exist; domain-aware system design and model adaptation create the path to close them.

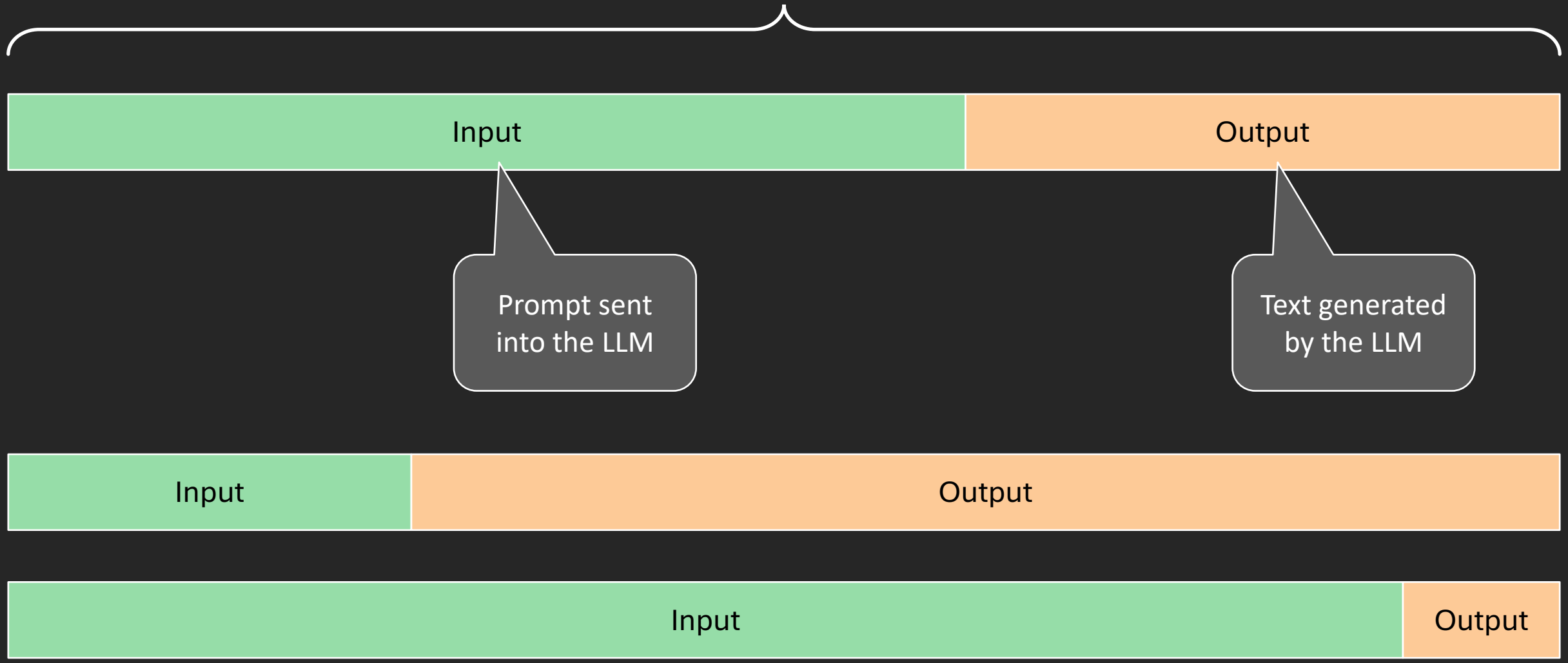
TOGETHER…SHAPING THE FUTURE OF ENERGY®

www.epri.com

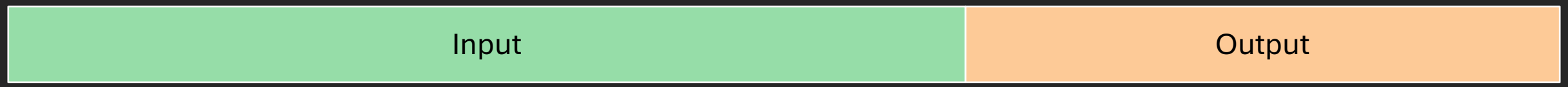
Flexibility is the foundation of modern energy reliability.



Context window (tokens)



Context window (tokens)



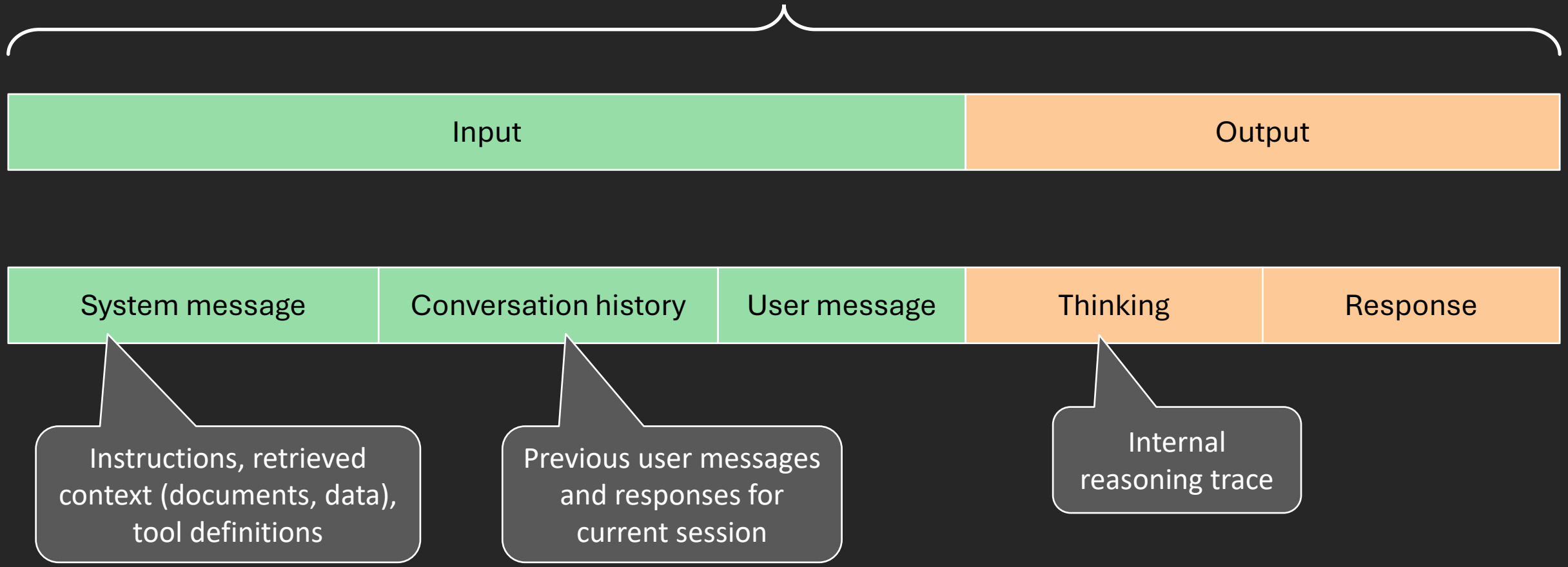
Typical maximum context length:

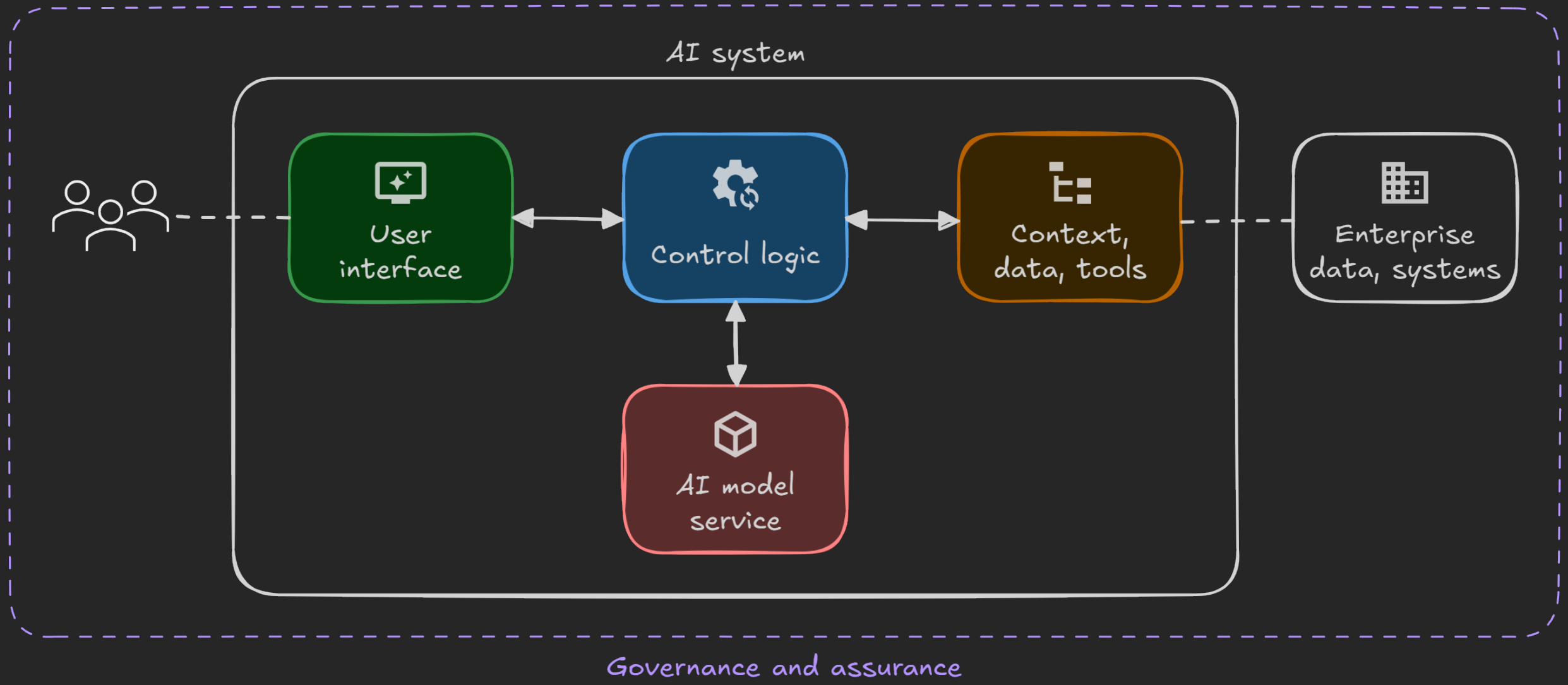
100,000–250,000 tokens

Some models support 1M+

Printed pages	Tokens
10	6,500
100	65,000
150	100,000

Context window

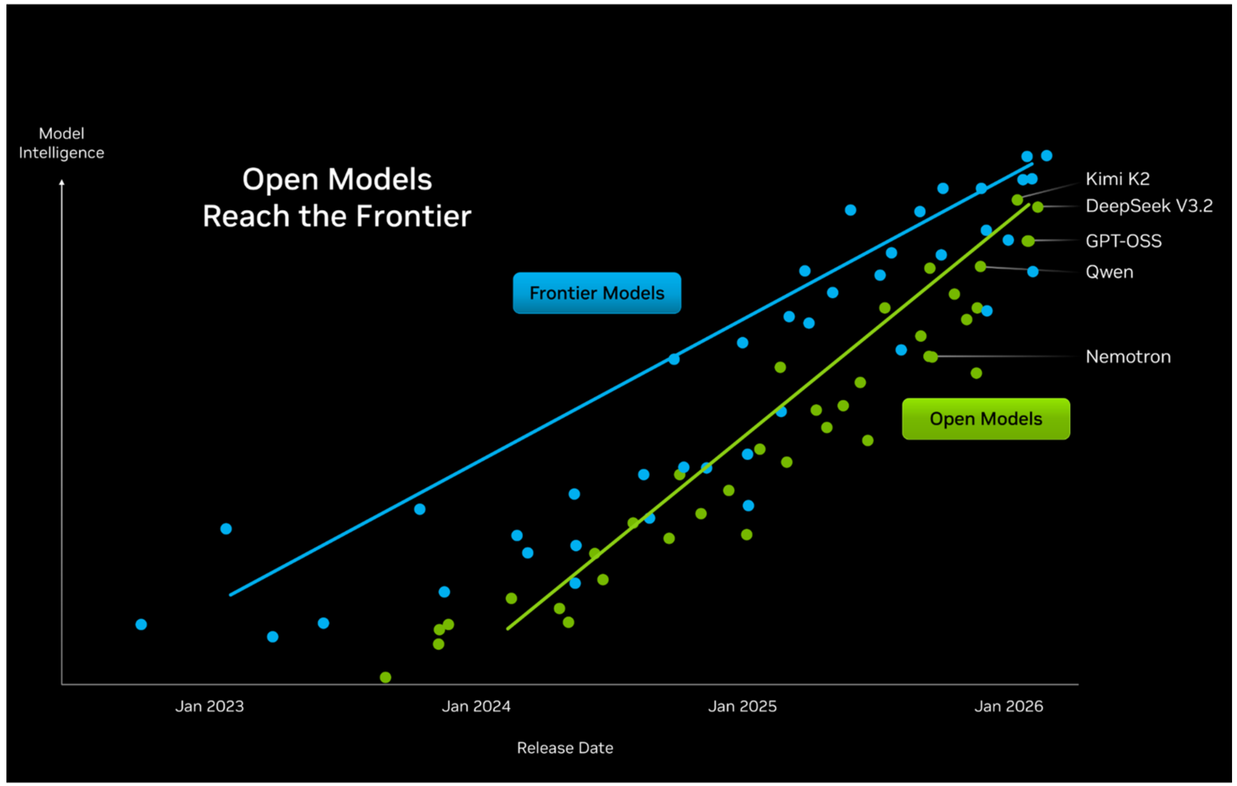


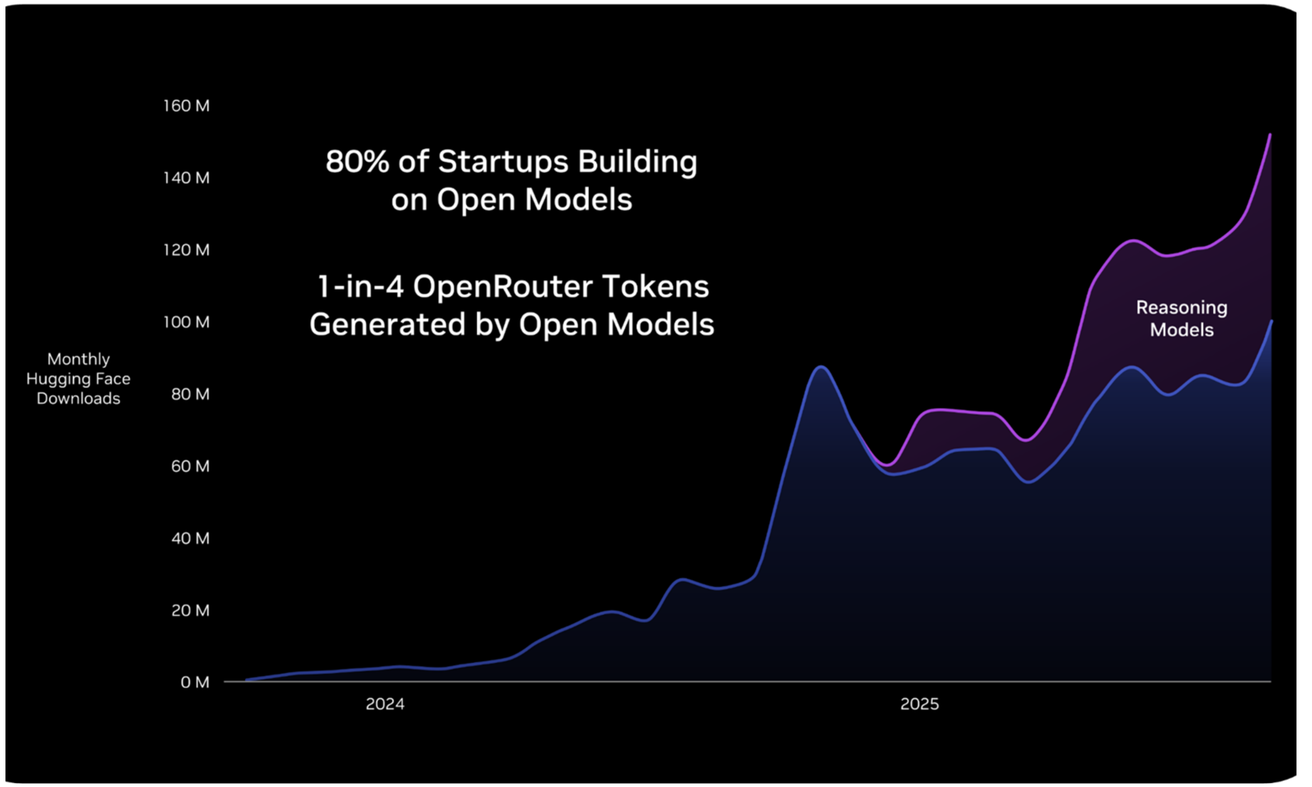


NVIDIA Perspective on Open Models for Specialized AI

Jihyun Yang | Solutions Architect | NVIDIA
jihyuny@nvidia.com

Open-Source AI is the Backbone of Innovation





NVIDIA

Why Open Models Matter for Technical Domains

Control, adaptation, and trust at domain scale

- Domain data is often private, regulated, or fast-changing
- Success is task-specific: valid outputs, correct tool use, grounded answers
- Open models give teams control over adaptation, deployment, and cost
- But openness alone is not enough—grounding and measurement still required

Data Control

Run models closer to your proprietary data

Task Fit

Adapt to domain terminology and success criteria

Cost + Speed

Optimize inference for high-volume agentic workflows

Trust + Audit

Inspect, validate, and govern model behavior

NVIDIA Strategy: Open + Efficient + Customizable

Three pillars that turn model weights into domain -ready AI systems

Open

Transparent model assets and developer ecosystem
Developers can inspect, adapt, and deploy with full visibility

Efficient

Practical inference for agentic and high-volume workloads
Every agentic workflow multiplies model calls—cost matters

Customizable

A repeatable path from baseline model to domain-specific system
Nemotron + NeMo + NIM + Retriever + Guardrails + Dynamo

NVIDIA Strategy: Open + Efficient + Customizable

Three pillars that turn model weights into domain -ready AI systems

Open

Transparent model assets and developer ecosystem
Developers can inspect, adapt, and deploy with full visibility

Efficient

Practical inference for agentic and high-volume workloads
Every agentic workflow multiplies model calls—cost matters

Customizable

A repeatable path from baseline model to domain-specific system
Nemotron + NeMo + NIM + Retriever + Guardrails + Dynamo

Nano
30B-A3B

Super
120B-A12B

Ultra
550B-A55B

NVIDIA Strategy: Open + Efficient + Customizable

Three pillars that turn model weights into domain -ready AI systems

Open

Transparent model assets and developer ecosystem
Developers can inspect, adapt, and deploy with full visibility

Efficient

Practical inference for agentic and high-volume workloads
Every agentic workflow multiplies model calls—cost matters

Customizable

A repeatable path from baseline model to domain-specific system
Nemotron + NeMo + NIM + Retriever + Guardrails + Dynamo

Nano

30B-A3B

Super

120B-A12B

Ultra

550B-A55B

Open Datasets

Pre-Training Tokens, Post-Training Samples, RL Tasks

NeMo Gym

RL Environments & Skills

Open Research

Papers, Samples

NVIDIA

The Making of Nemotron 3 Super

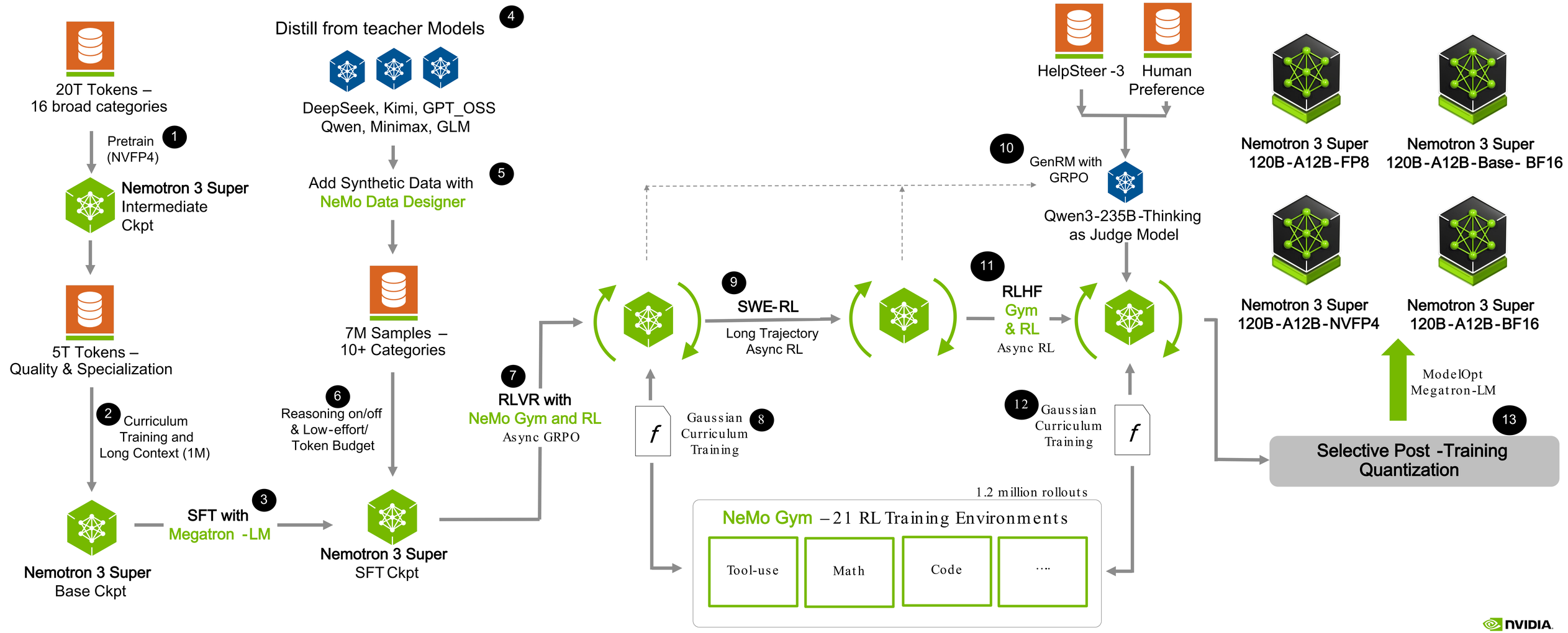
Latent MoE Hybrid Mamba –Transformer; Multi -token prediction; Multi -environment RLVR

Pre-training
25T tokens

Supervised Fine-Tuning
Agentic Skills and Reasoning

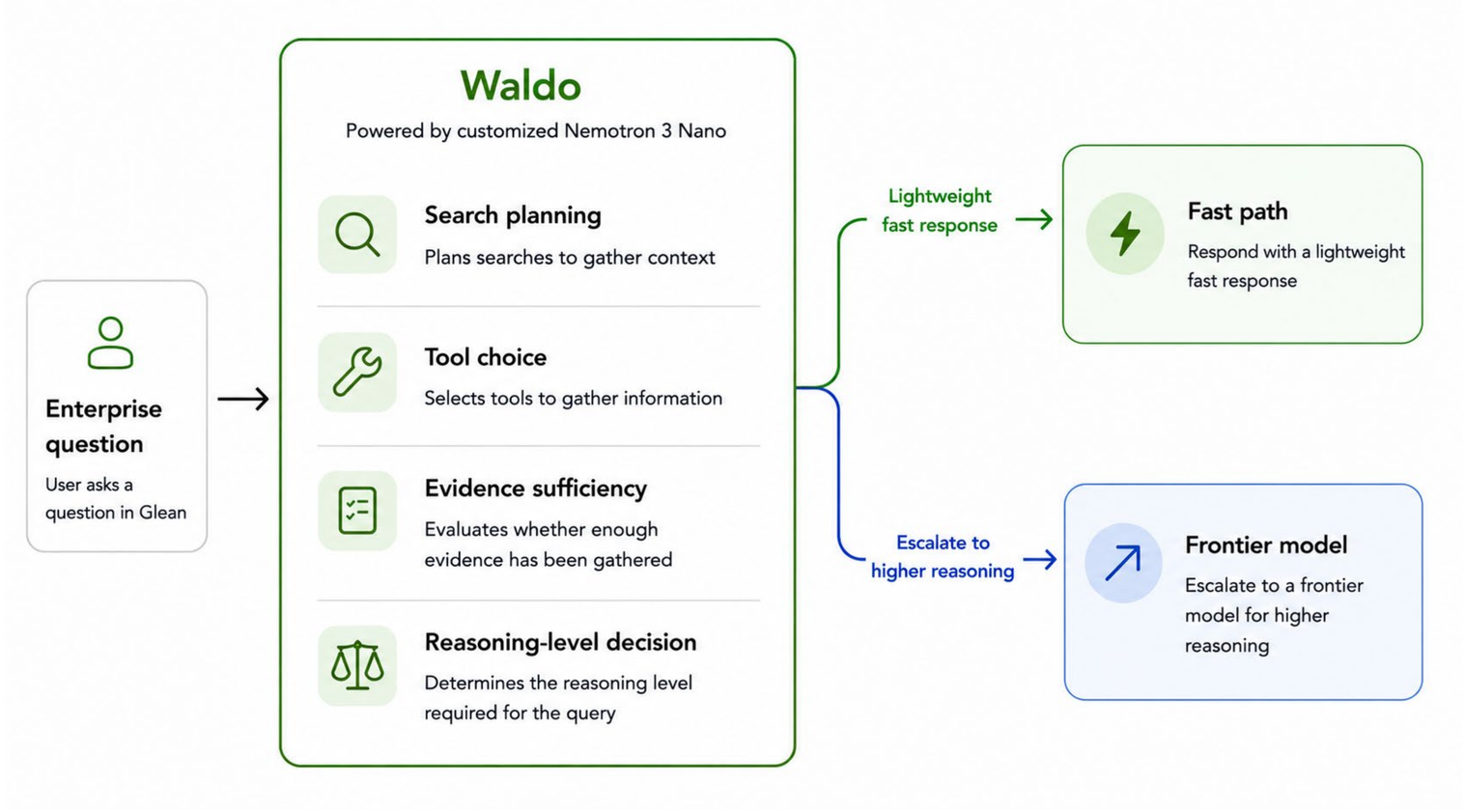
Reinforcement Learning (RL)
Multi -environment RLVR and RLHF with NeMo Gym and NeMo RL

Post-Training Quantization
Efficient Inference



Glean – Balancing Performance and Latency with Nemotron

Waldo: Agentic Search Model Built on Nemotron 3 Nano



50%

25%

0%

Agentic AI is Built on a System of Models

Agents often combine proprietary and open models

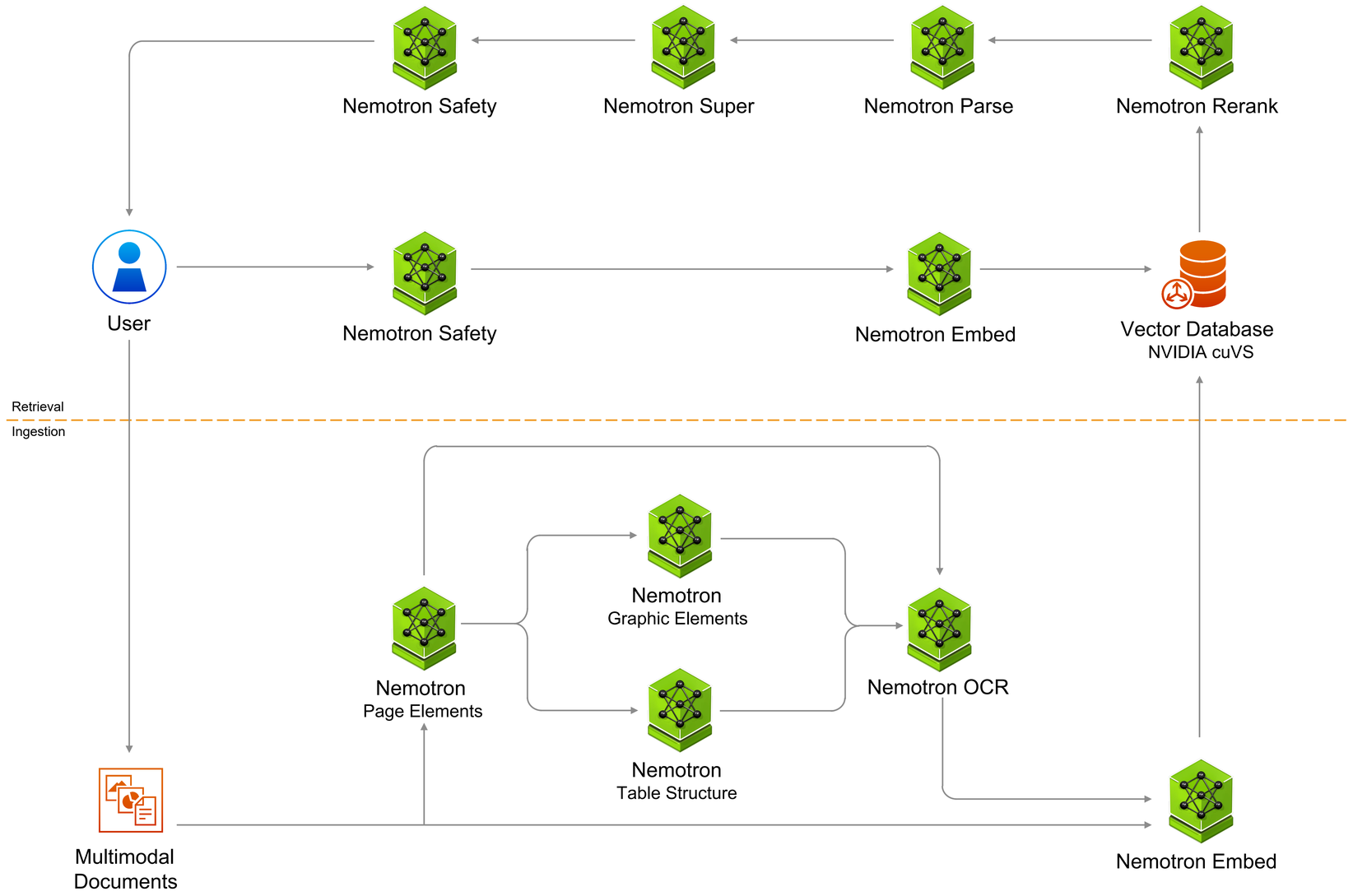
Practical Adoption Pattern

Start small and measurable —measure first, customize only where evals show a gap

1	**Pick one workflow**	Choose a specific workflow with a clear business or technical outcome
2	**Build eval set**	Create a small, representative evaluation set that captures the real task
3	**Baseline + ground**	Compare open model baseline against a grounded RAG or tool-using system
4	**Customize**	Add synthetic or curated data only where evals show a gap; apply SFT/LoRA or RL
5	**Deploy + improve**	Deploy with NIM and guardrails, monitor in production, feed signals back into eval

Retrieval -Augmented Generation Blueprint

Connect AI to Your Data: Accurate Answers Grounded in Organizational Knowledge




NVIDIA