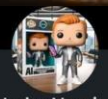
KAMIWAZA

Superhuman Power for Enterprise AI



Luke Norris
Wearer of white shoes / Builder of companies that make an impact

KAMIWAZA

KAMIWAZA Intro:

At **KAMIWAZA**, our mission is to empower enterprises for the 5th industrial revolution, aiming to be the cornerstone of technology as they scale to 1 trillion inferences per day and beyond.

KAMIWAZA : Like every other fascinating foreign word, **KAMIWAZA** is hard to translate. The closest is ''**technique' (waza) of the 'god' (kami**)'. **KAMIWAZA** is where you try to be the myths, the superheroes; therefore we like to paraphrase Seth Godin in the Icarus Deception: It's when you try to fly closer to the sun and become superhuman.

Thus **KAMIWAZA** is pursuit of Superhuman capabilities for the Enterprise.

Founded in 2023 by CEO Luke Norris and CTO Matt Wallace notably former CEO/CTO of Faction Inc a Multi-Cloud Data Service focused on the Global 2000 market segment.

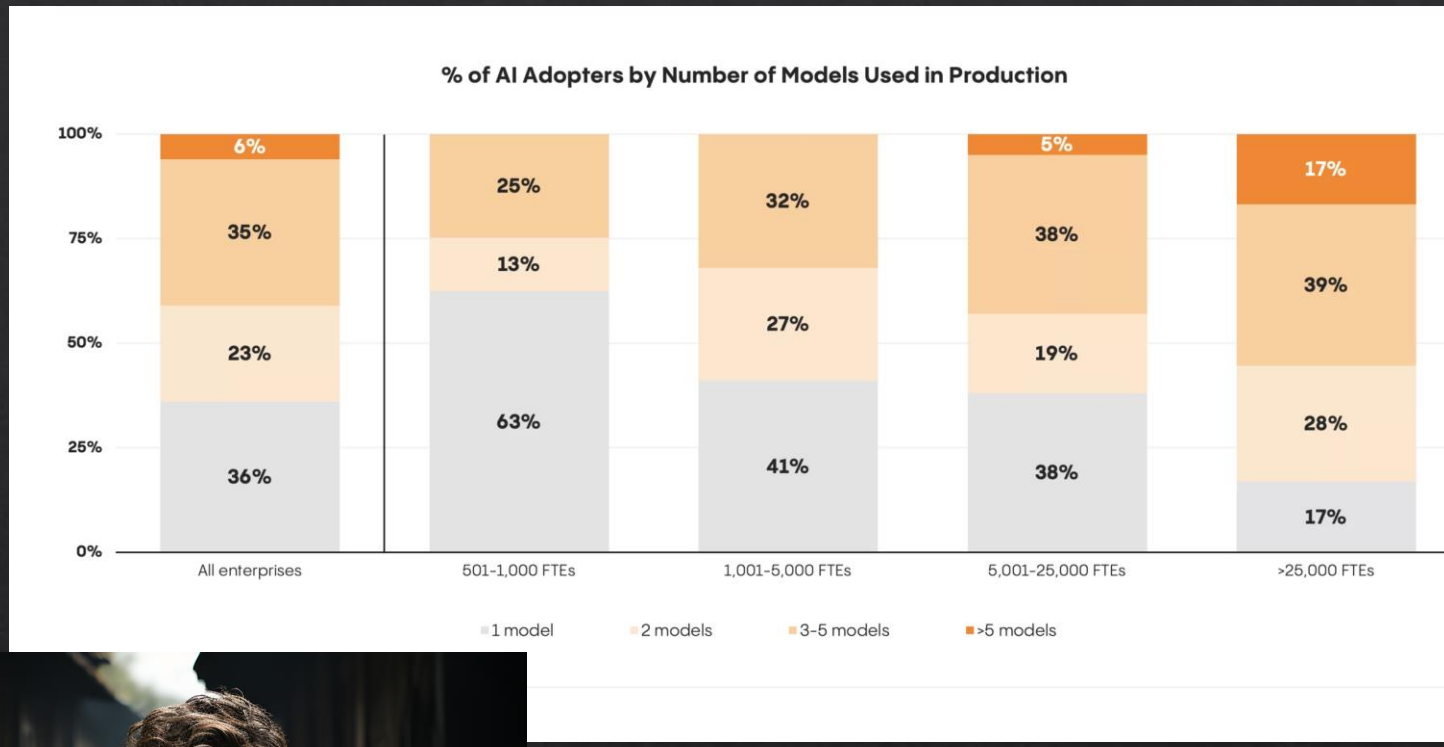
The Fifth Industrial Revolution #5IR

The 5th Industrial Revolution, often referred to as Industry 5.0, represents a transformative era where advanced technologies, particularly artificial intelligence (AI) and human creativity, converge to create highly interconnected and intelligent systems. Unlike the 4th Industrial Revolution, which focused on automation and the integration of cyber-physical systems, the 5th Industrial Revolution emphasizes the collaboration between humans and machines. This partnership aims to enhance productivity, innovation, and efficiency while maintaining a strong focus on sustainability and human-centric solutions.

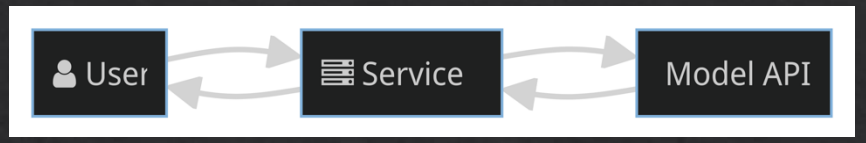
Enterprises are NOT training models

Enterprises are using and enhancing multiple Foundation Models 6+.

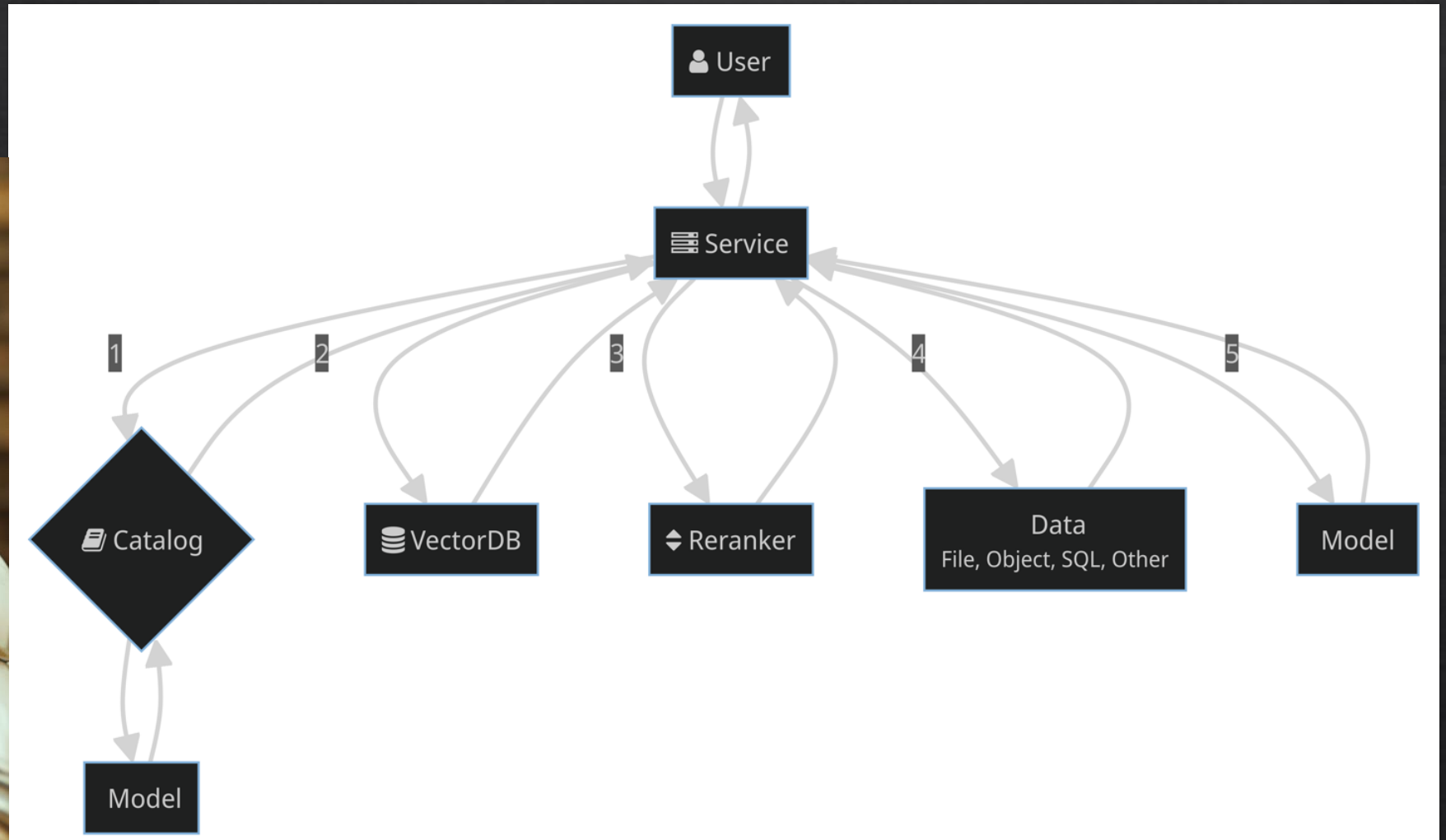
Worldwide, there are 30 million developers, 300,000 ML engineers, and only 30,000 ML researchers. For those innovating at the very forefront of ML, our references estimate there may only be 50 researchers in the world that know how to build a GPT-4 level system.



Chat - 2 or 3 Inferences per Request

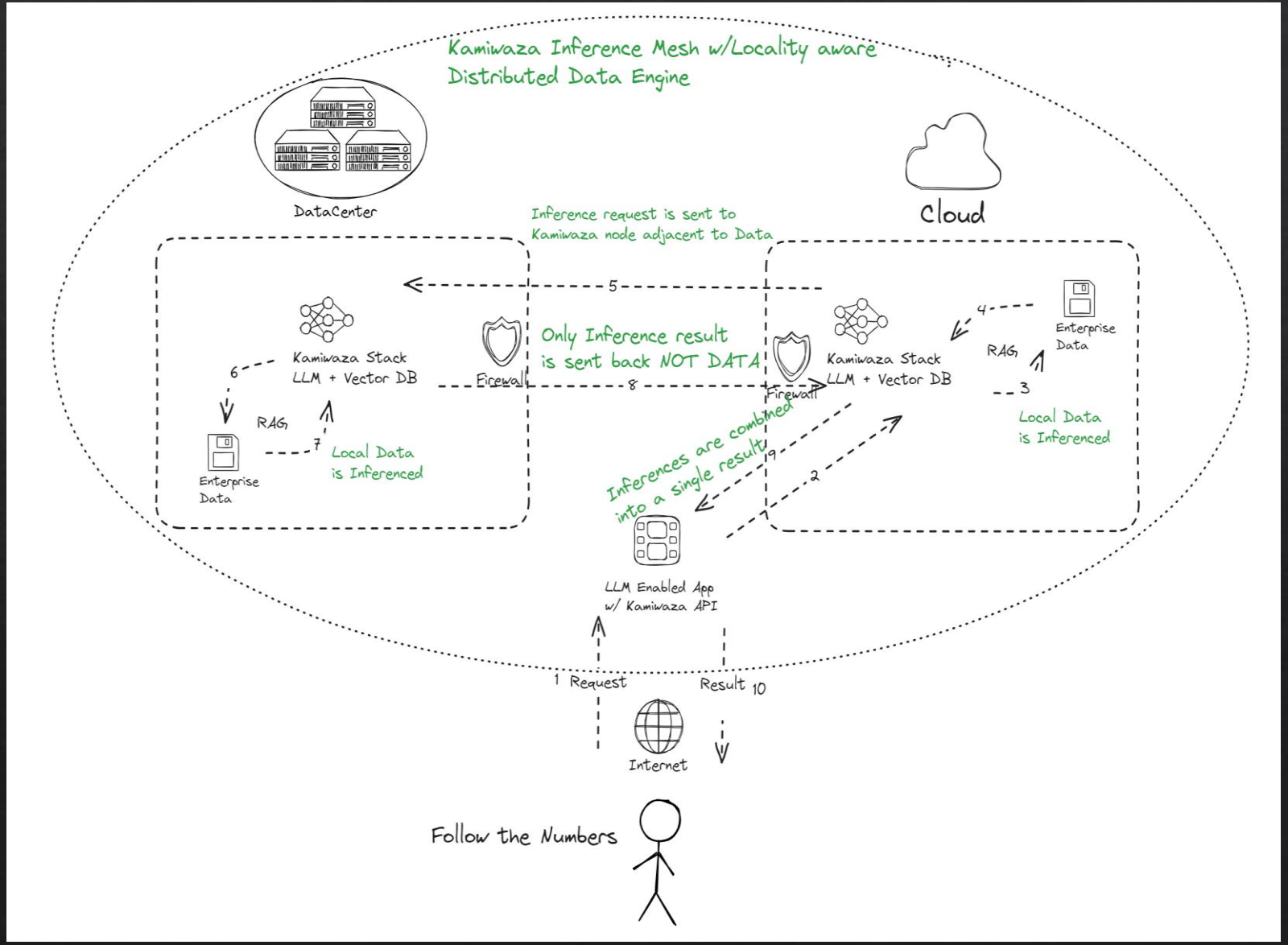


RAG - Retrieval Augmented Generation 5-30 Inferences per Request : Private Data – Lingua Franca of Enterprise



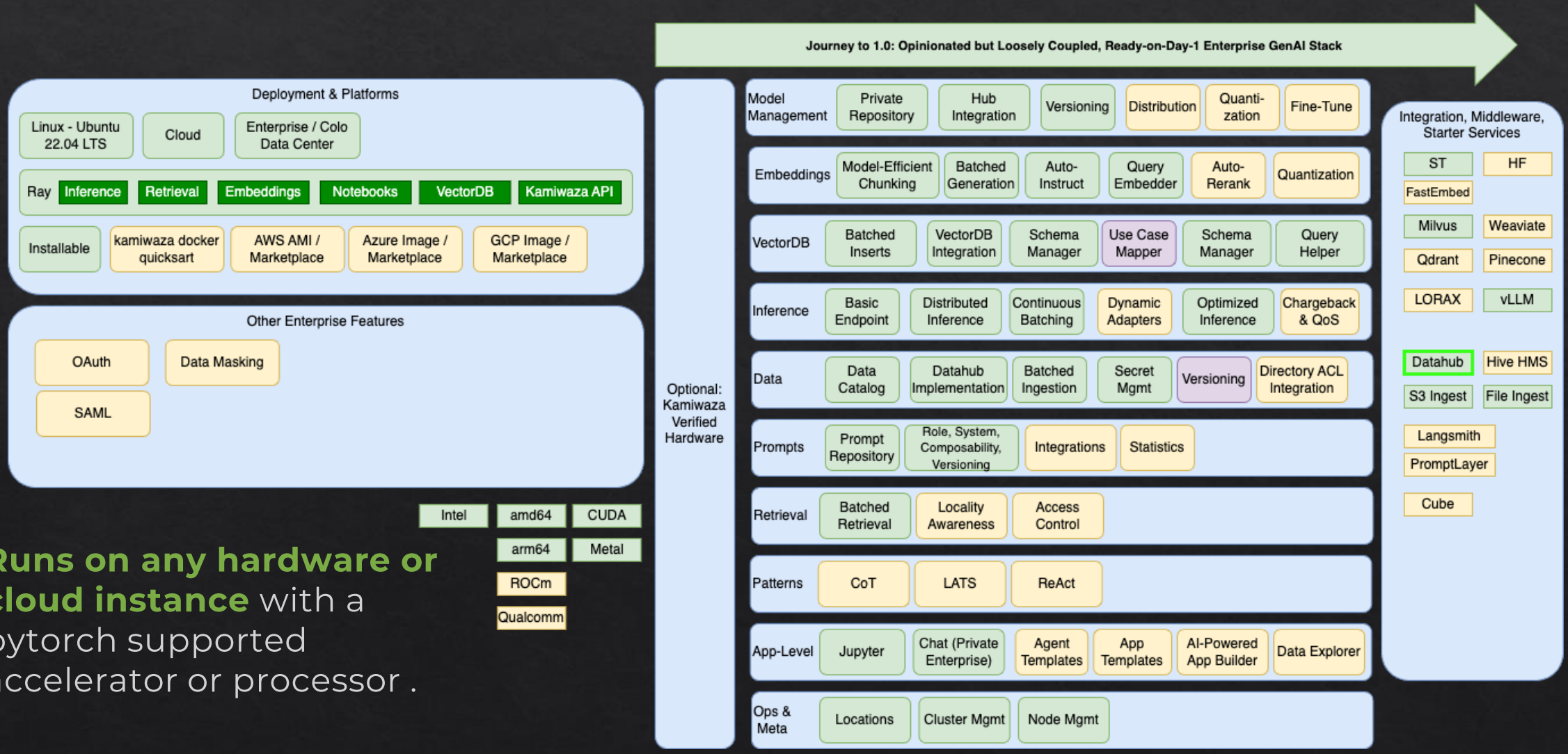
KAMIWAZA GenAI Engine Enables Enterprise AI Anywhere

KAMIWAZA's GenAI stack focuses on two novel technologies to enable PRIVATE Enterprise AI Anywhere, **Inference Mesh** and **Distributed Data Engine. These two in combination** provide locality-aware data for RAG capable of inference processing where the data lives regardless of location, across **on-prem, cloud, edge.**



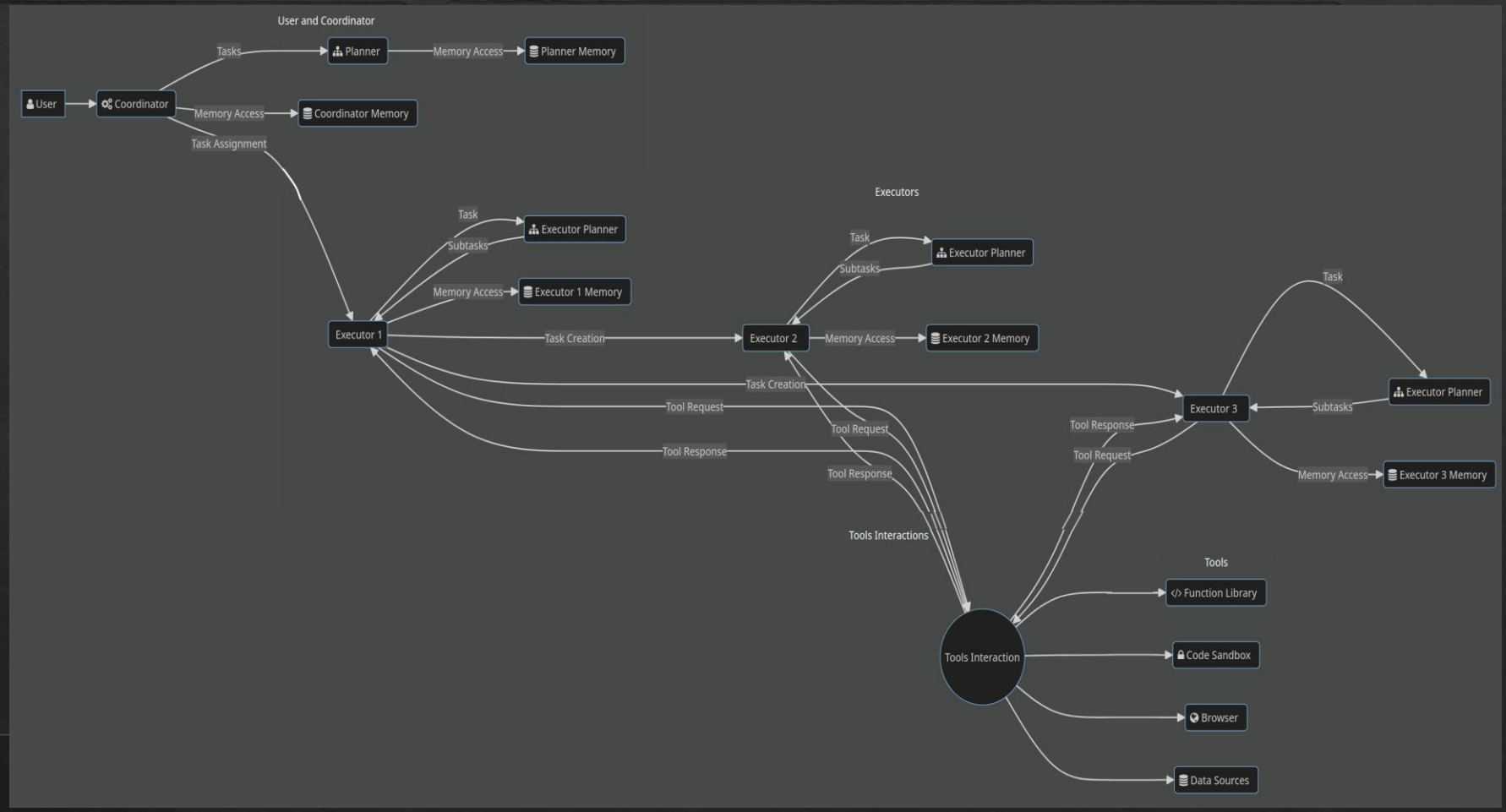
KAMIWAZA Stack V1

Same developer experience from a laptop to N-scale distributed retrieval and inference mesh, with private model management, pre-integrated with best of breed tools, in an **opinionated** but loosely coupled stack.



Runs on any hardware or cloud instance with a pytorch supported accelerator or processor .

Agents – 100s – 1,000s-10,000s of Inferences per Action "this is Autonomous and Non-Human scale”



2019 -- https://openai.com/index/emergent-tool-use/

"We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run“

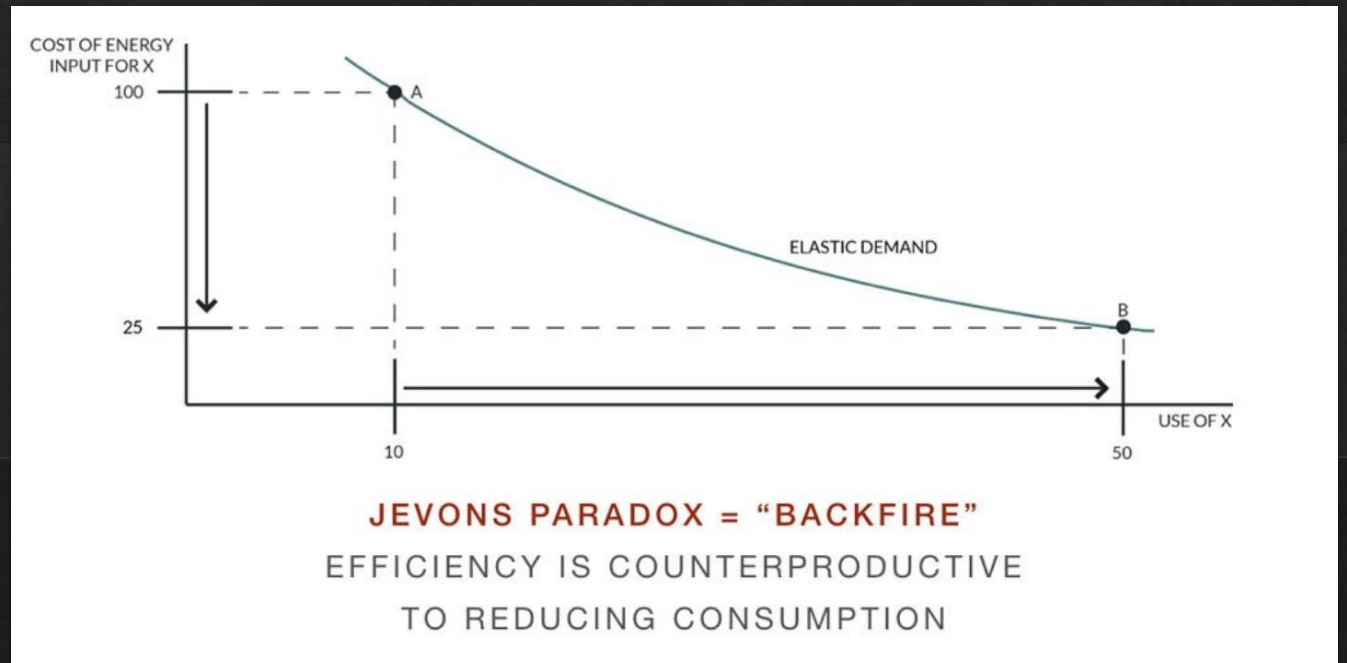
This highlights the tendency of people to be overly optimistic about the immediate impact of new technologies while failing to foresee their long-term potential and transformative effects.

Roy Amara – Amara’s Law

Jevons Paradox for the 5th Industrial Revolution

- During the 19th century, Britain experienced rapid industrial growth, heavily dependent on coal as the primary energy source. Coal powered factories, transportation, and homes, driving economic expansion.

- Jevons concluded that the increase in coal efficiency led to an increase in the total consumption of coal rather than a decrease. He argued that improvements in resource efficiency could lead to higher overall resource use due to increased economic activity and demand.



"As a result of these reassessments of technology capabilities due to generative AI, the total percentage of hours that could theoretically be automated by integrating technologies that exist today has increased from about 50 percent to 60–70 percent. The technical potential curve is quite steep because of the acceleration in generative AI's natural-language capabilities."

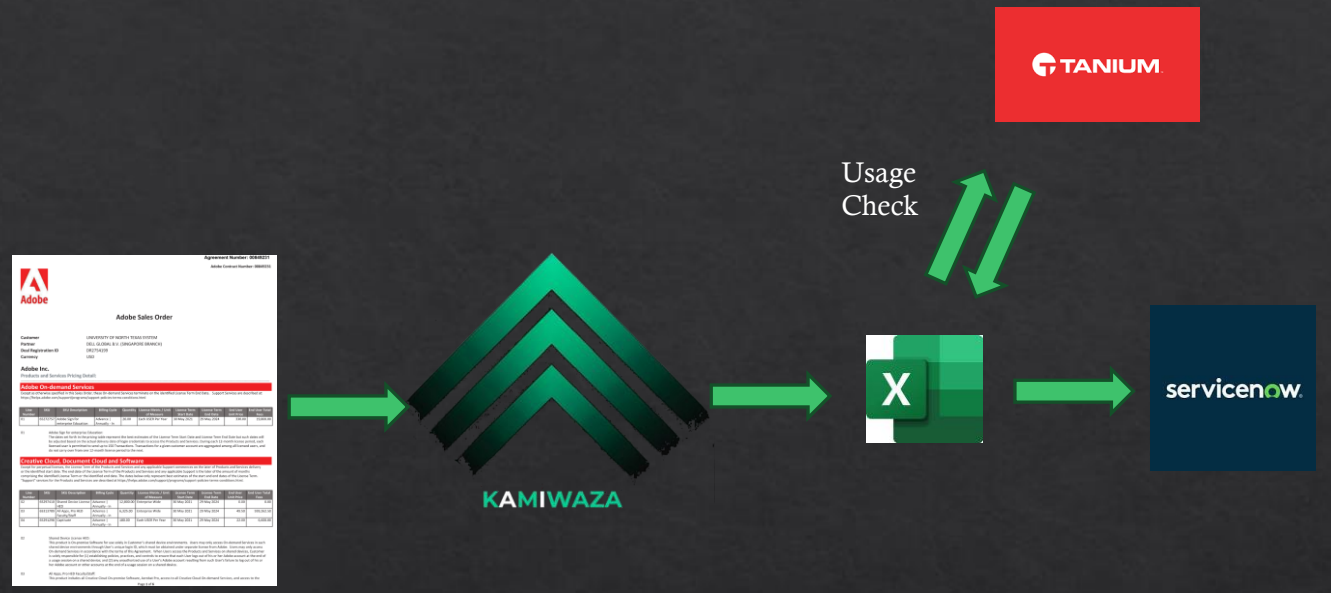
https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

- Access File store of Contracts
- Ingest Contracts (OCR/LLAVA)
- Find Entitlements
- Sanity Check / Guard Rail
- Output CSV of Entitlements
- Import to Service Now
- Verify from Sanity Check

- Simple non-human scale workflow –

Automate contract and entitlement review via AGENT workflow

- Access File store of Contracts
- Ingest Contracts (OCR/LLAVA)
- Find Entitlements
- Sanity Check / Guard Rail
- Output CSV of Entitlements
- Import to Service Now
- Verify from Sanity Check
- check Usage with Tanium
- Update Service Now



- Simple non-human scale workflow expanded –

Automate contract and entitlement review via AGENT workflow

Enhancing ML/Deterministic workflow with GenAI

Data Producer Application

↓

Data Ingestion

↓

Initial Data Validation & Quality Check

(Deterministic Rules & AI Agent)

↓

Hybrid Data Transformation

(Deterministic + AI-Driven Transformations)

↓

Data Annotation & Enrichment

↓

Structured Data Formatting

(Deterministic + AI Validation)

↓

Data Storage & Versioning

↓

Intelligent Data Distribution

(Deterministic + AI Optimization)

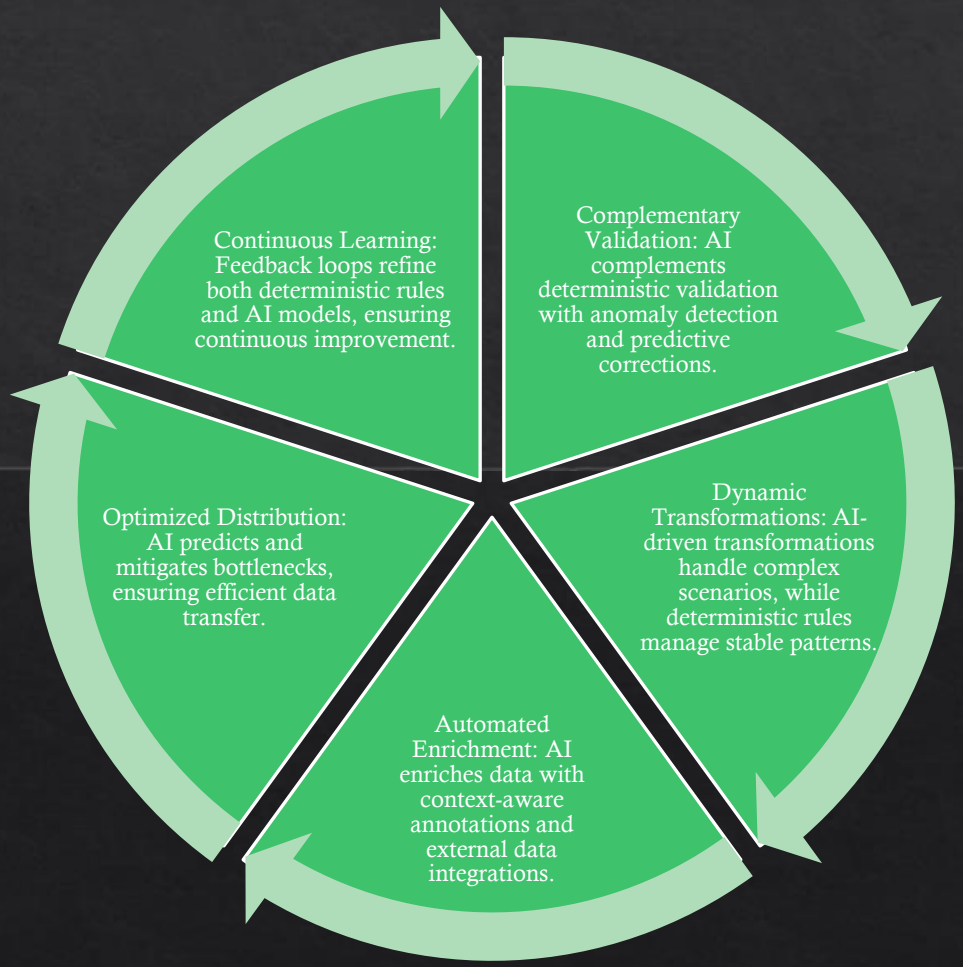
↓

Consumer Application Utilization

↓

Continuous Monitoring & Feedback Integration

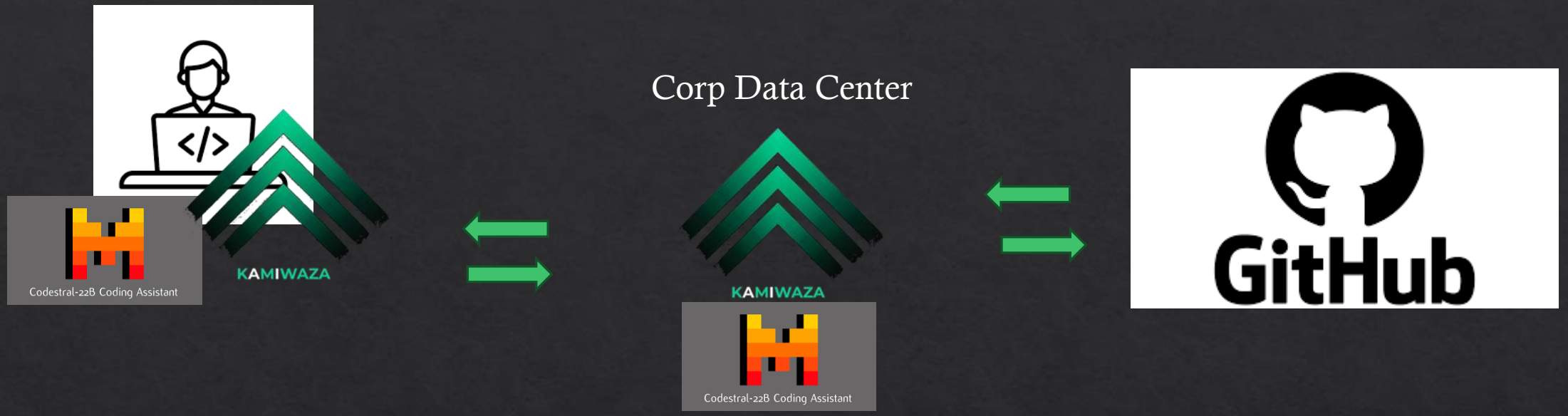
↺



KAMIWAZA

- **Data Collection**
- IoT Sensors → Central Database
- **Data Processing**
- Data Cleaning → Data Analysis → Metric Calculation
- **Reporting and Compliance**
- Tax Calculation → Report Generation → Compliance Check
- **Submission and Archiving**
- Automated Submission → Secure Record Keeping
- **Continuous Improvement**
- Feedback Loop → Process Adjustments

Automate Tax utilization reporting for OIL and Gas Wells



Enhance coders with the latest code assistant while automating code checks validations and regression testing via Agents – Balancing Inference on Laptop running models locally along with Models in Data Center Agents.

KAMIWAZA resolves the direct pain for Enterprise AI

Interviewed Executives of F500 on issues preventing or slowing AI adoption

Multiple Unique disparate stacks Cloud, Core and Edge.

- Single Stack deployed as an installable package or containers for Core/Edge, or a cloud image in your cloud of choice
- Consistent experience from developer laptop to production

Private AI with Private Data is a must

- **Use with existing Open-Source models**
- **Fine tuning for Enterprise requirements (Knowledge, Tone, Use Case)**

Data Gravity across Cloud/s, Core and Edge

- **Internal catalog with discovery**, metadata, and ingestion tools; integration with external catalogs such as Hive Metastore, Unity Catalog; helps create **Location-Aware** data and inference mesh
- Automated workload distribution can scale up and down (including to zero) in response to application demand

Inference at Scale of Production

- Utilizes Ray to provide a distributed inference framework that intelligently scales across enterprise hardware or private cloud resources
- **Inference Mesh seamless integrates stacks and data across multiple clouds, core and edge**

Blue Green Deployments

- Artifactory-style local model and metadata repository
- Facilitates regression tests allowing you to validate iterations of prompts, models, and stack versions

System Integration and Identity Access

- **OATH and SAML**
- **Secret key management integrated with data retrieval across Cloud, Core and Edge**

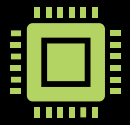
Need an outcome without internal Experts

- **Turnkey Opinionated stack deployable via Docker Quick Start**
- Point to over 40 types of Data sources and beginning the full RAG pipeline to educate any model from Hugging face and other repositories
- Repository of application patterns designed for enterprise use, helping as a “jump start” catalog for more complex activities like Chain of Thought, Chain of Density, ReAct tool indicators, etc.

What hardware to buy where to start

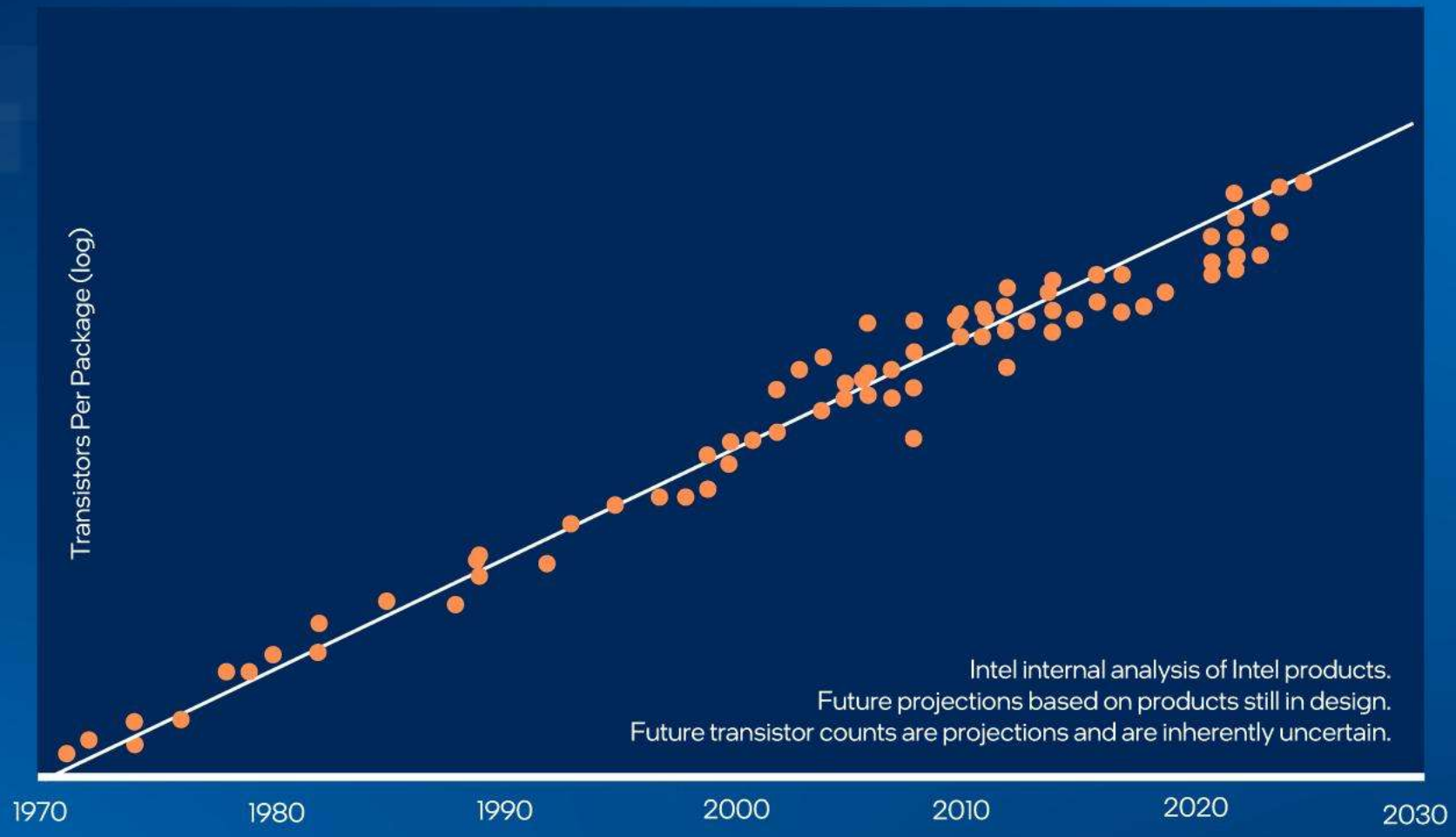
- **Validated Design with key vendors tied to use cases**
- **Integrations with Enterprise and Cloud storage**

Knowledge work is now Tokens !



Knowledge work is done nearly on Silicon – costs is governed by things such as Moore’s Law

Law of Accelerating Returns — the tendency for advances to feed on themselves, increasing the rate of further advance, and pushing well past what one might sensibly project by linear extrapolation of current progress.



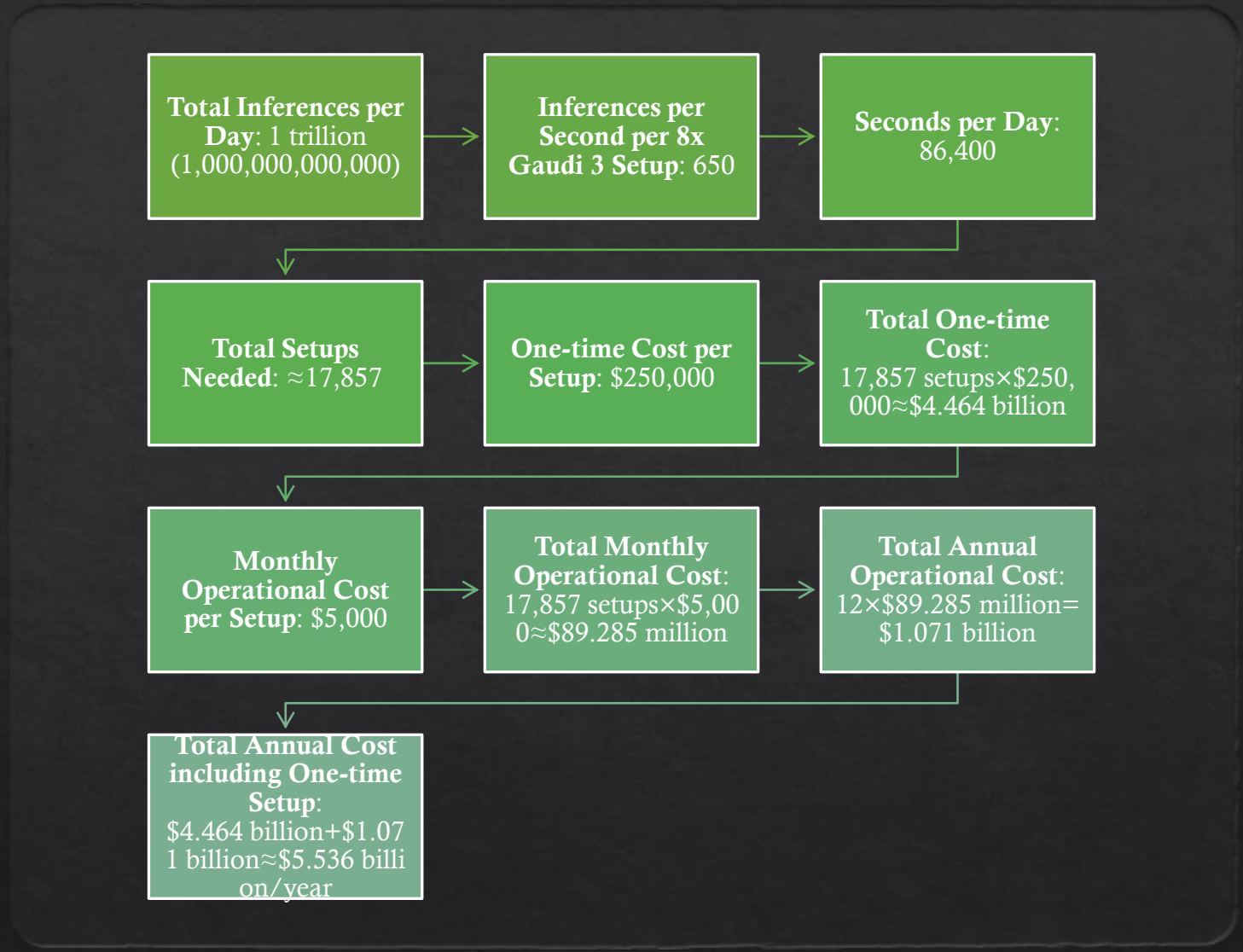
Aspiring to 1 Trillion transistors in 2030

- ✓ RibbonFET
- ✓ PowerVia
- ✓ High NA
- ✓ 2.5D/3D packaging

intel

1 Trillion Inferences - aaS -

1 Trillion Inferences with Gaudi 3



Impact for $5.5B a year aka 1Trillion Inferences a Day

If we suppose a knowledge worker represents about 2500 workflows a day 1 Trillion Inferences would support 80,800 knowledge workers. At an average of $115,000 costs per worker per year that's 9.3B of value on ~4.4B one time cost and 1.1B operating Annual costs.

Costs Feasible today at any scale!

Costs Decreasing 1000x in 1-2 year via Multiplicative impacts in Inference market especially with the impact of the UXL and OneAPI standards

4 Archetype types of enterprise Inferencing

•**Discrete Accelerator Card-Based Servers**: Cost-effective and scalable, ideal for initial deployments with

modest needs, offering flexibility to grow as requirements increase.

•**Large-Format Specific GenAI Accelerator Systems**: High-performance for critical, high-throughput

applications, suitable for enterprises with established AI workloads.

•**On-Device AI Processing**: Leverages existing devices to enhance responsiveness and reduce central

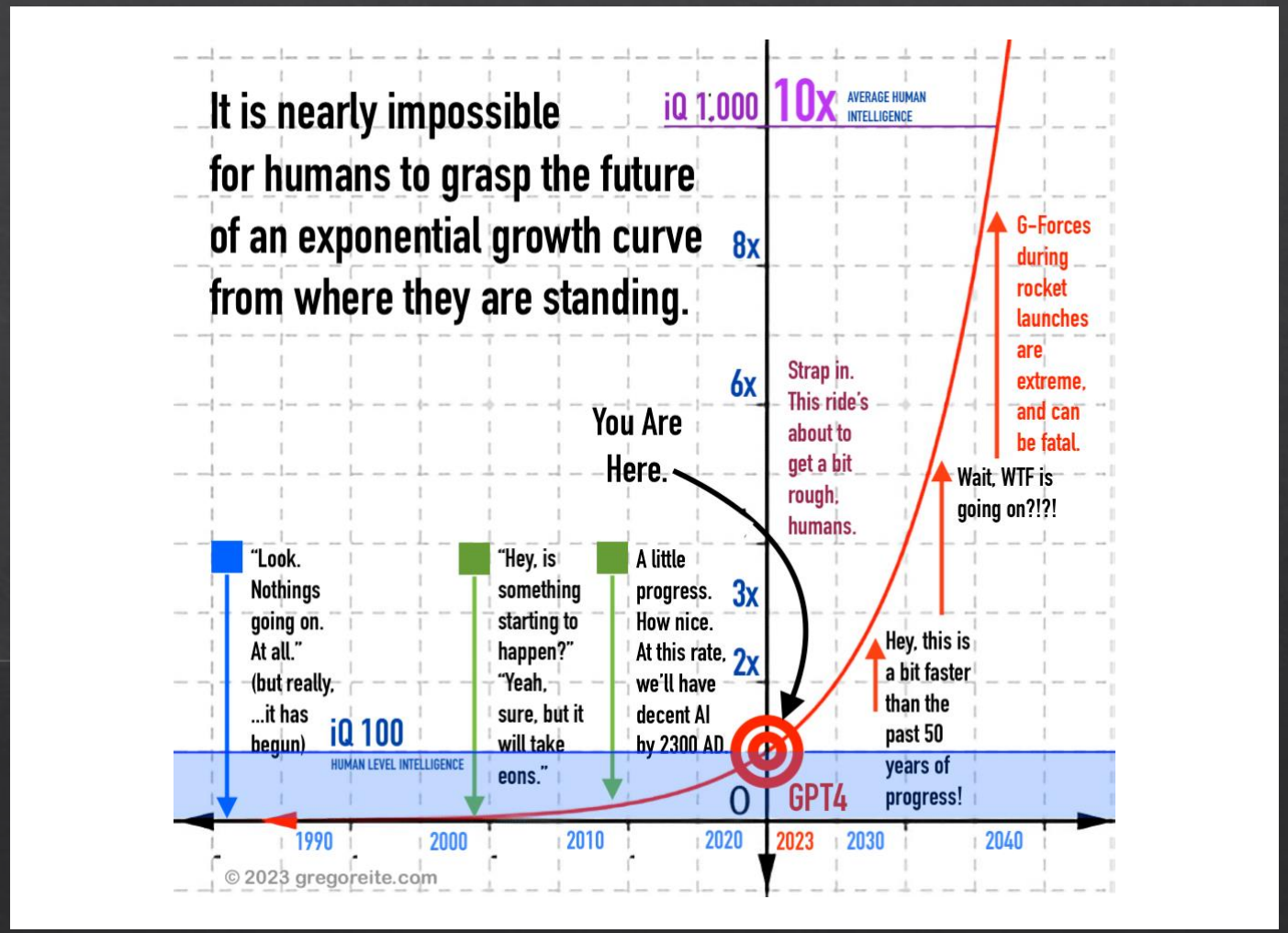
infrastructure demands, ideal for distributed processing.

•**LLM-Specific Processor Clusters and AI ASICs**: Emerging high-performance solutions for large-scale AI

workload consolidation, requiring significant investment.

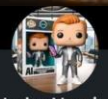
Enterprise Take-off is analogous to AGI Take-off once it starts you can't catch up

- If a 5IR Enterprise at 25% of workflow
- Ai improves Ai
- "Ai Kaizen" 1% efficiency a week or month
  - Weekly Improvement Impact: The workflow will be 41.9% efficient due to automation at the end of the year.
  - Monthly Improvement Impact: The workflow will be 28.2% efficient due to automation at the end of the year.



KAMIWAZA

Superhuman Power for Enterprise AI



Luke Norris
Wearer of white shoes / Builder of companies that make an impact