Building a Generative AI Flywheel for the Organization

Presented by: Brian Sathianathan, Co-Founder and CTO/CDO, Iterate.ai & Kevin Homer, VP of Sales, Iterate.ai

About this talk

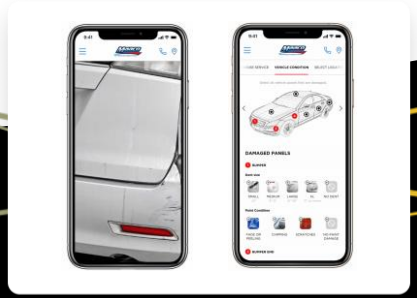
This exclusive webinar showcases Generate powered by Iterate.ai, a Personal AI Assistant that uses Retrieval-Augmented Generation (RAG) and multiple secure vector databases to provide LLM capabilities to the Enterprise and SMB markets.

Unlike other AI Assistants, Generate is deployed locally. This Private LLM not only allows users to leverage the application offline, but it removes the risk of exposing private company data to the outside world when connecting to public-facing chatbots.

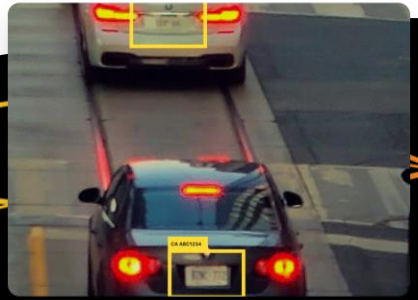
When it comes to deployment, Generate is extremely flexible. The Enterprise version can be deployed as a Docker instance in a private or public cloud, or shipped as a ready-to-install appliance. Generate can also be deployed on an AI PC (Intel Core Ultra Processor 155U and above).

Iterate is a complete innovation platform

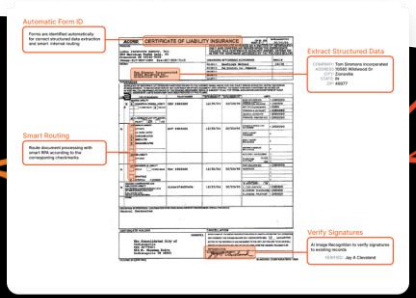
GenAI for LLM, Voice, and Images



Intelligent Applications (AI/AR)



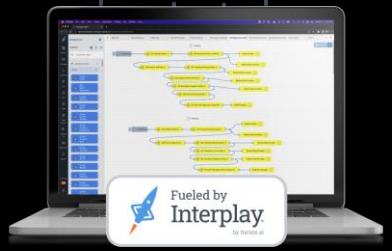
Computer Vision on Edge



Intelligent Document Processing

4,100
AI Edge Deployments

40MM
Monthly Users



6MM
Monthly Licence Plates

15K
Daily Documents

About Iterate.ai

Recognition and Partners

- **Best Workplaces for Innovators**, **Artificial Intelligence** and Robotics, Fast Company, 2024 (*Sept reveal, under embargo)*
- **AI 100** (world's top AI firms), Knowledge Management World, 2024
- Company of the Year, Colorado Tech Assoc, 2023
- Inc. Magazine's Best-in-Business: **AI** and **Data**, 2023
- **AI 100** (world's top AI firms), Knowledge Management World, 2023
- BIG Innovation Award 2023
- Long-Term Partner of the Year with Ulta Beauty, National Retail Federation, 2022
- World's Best Low-Code/No-Code Platform Finalist, Software Internet Industry Association (SiiA), 2022
- Colorado's **Top AI** Company, 2022
- USA's Top Technology Innovation, DSA USA, 2022***
- Industry Innovation of the Year Canada, 2022**
- Intellyx Analyst's Digital Innovator Award, 2022
- Colorado Company to Watch, 2022
- Deloitte Fast 500: Grow 284% over 3 years, 2021
- Top Convenience Store Innovation Europe, 2021*
- World's Best Low-Code/No-Code Platform Finalist, SiiA, 2021

* Circle K / Couche Tard, ** Pampered Chef, Berkshire Hathaway, *** Direct Selling Association

Resellers and Partners

FUJIFILM, HUGHES An EchoStar Company	Reselling Iterate.ai Apps plus Edge Compute
intel, Qualcomm	Partnered to Optimize LLM processing on CPUs for LLMs
aws, Google Cloud Partner	Marketplace + Extensive GPU/ML Support
NVIDIA, Azure	Hardware and Cloud Partners
Lenovo, DELL	OEM and hardware distribution
TD SYNNEX, World Wide Technology	System Integrators and resellers

Sample Clients

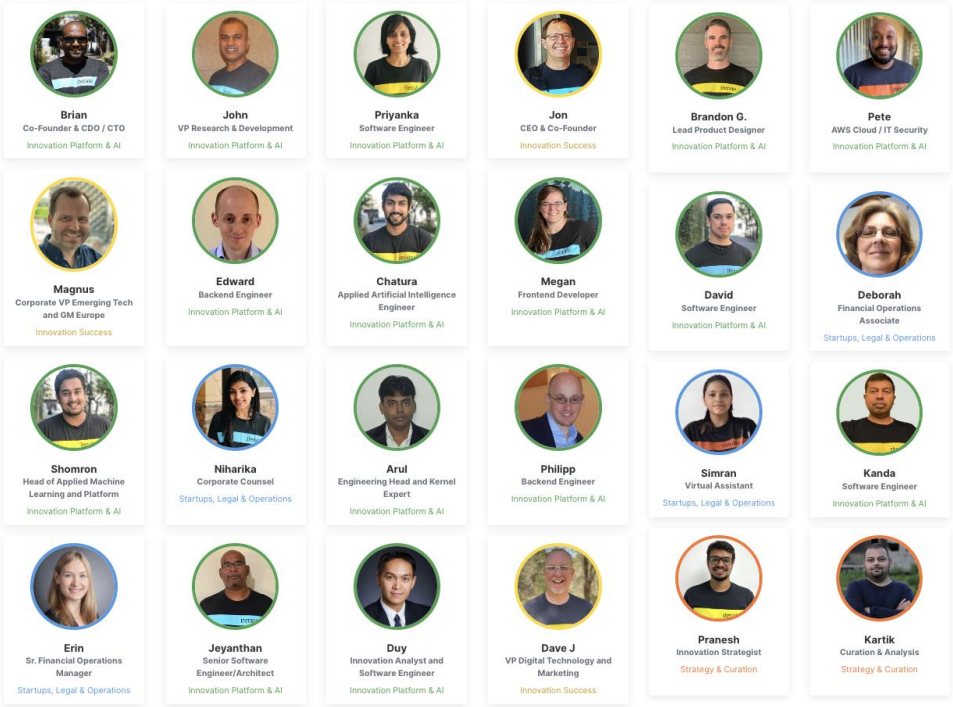
ULTA BEAUTY, FARRER PARK MEDICAL CENTRE, e.l.f. eyes lips face, FULLERTON HEALTH, STARBUCKS COFFEE, CHANEL, Couche-Tard., CIRCLE K, Great Eastern A member of the OCBC Group, UOB, pampered chef

Qualified Team

102
Iterators Globally

85%
Technical Background

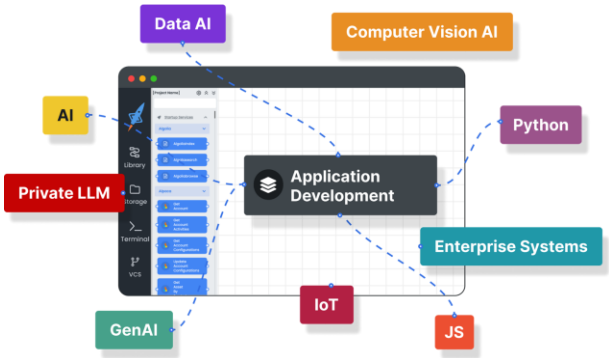
- ✓ 15% M.S., PhD in AI
- ✓ Kernel-Code
- ✓ AI-First Developers
- ✓ CX-Architects



Iterate Diversified Ecosystem

Interplay App Platform

- ➔ Low-code development studio
- ➔ Efficient runtime environment
- ➔ 4000+ nodes for integration
- ➔ Rapid prototyping
- ➔ Production-ready
- ➔ Enterprise level



AI Applications

- ➔ AI for the enterprise
- ➔ Generative AI, computer vision, staff management, dynamic data
- ➔ Finance, Retail, QSR, Automotive
- ➔ Secure, Private, and Flexible

Generate by Iterate.ai

Extract By Iterate.ai

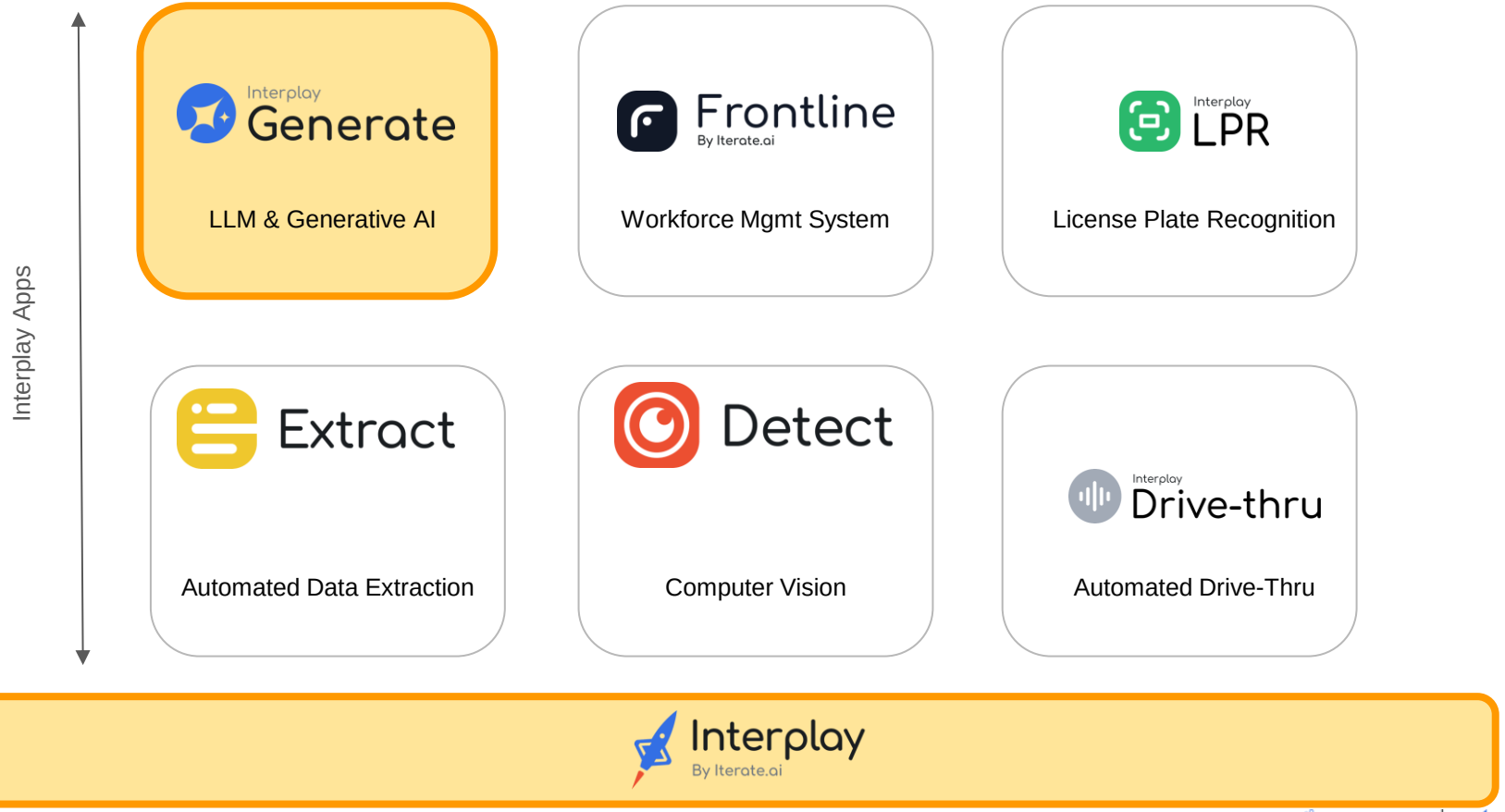
Detect by Iterate.ai

Frontline by Iterate.ai

Related Services (as needed)

- ➔ AI Strategy
- ➔ Model training
- ➔ Innovation Consulting
- ➔ Implementation Support

Interplay Platform and Ecosystem



Interplay: Low-Code AI Application

Interplay is a **complete platform solution**.

It **running in production** as middleware and finished solutions, plus it's a low-code **application development studio**.

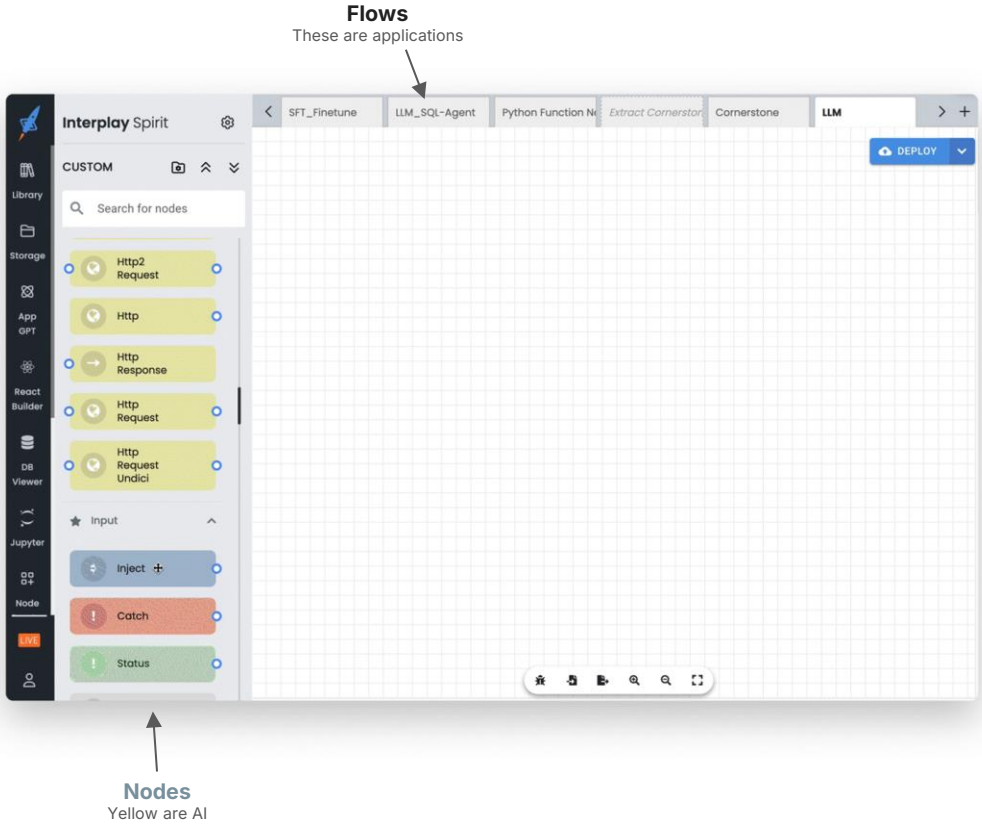
The studio includes several thousand pre-built **"nodes"** for AI, GenAI, databases, IoT connectors, communications protocols, and 3rd party SaaS APIs. About **200 are AI specific**. **Code can be modified** using node.js or python.

Interplay is running in a Docker container with **optimized runtime on the existing infrastructure**. It supports multi-cloud as well and on-prem or even embedded deployments. This **reduces vendor lock in** and **provides flexibility** for the future.

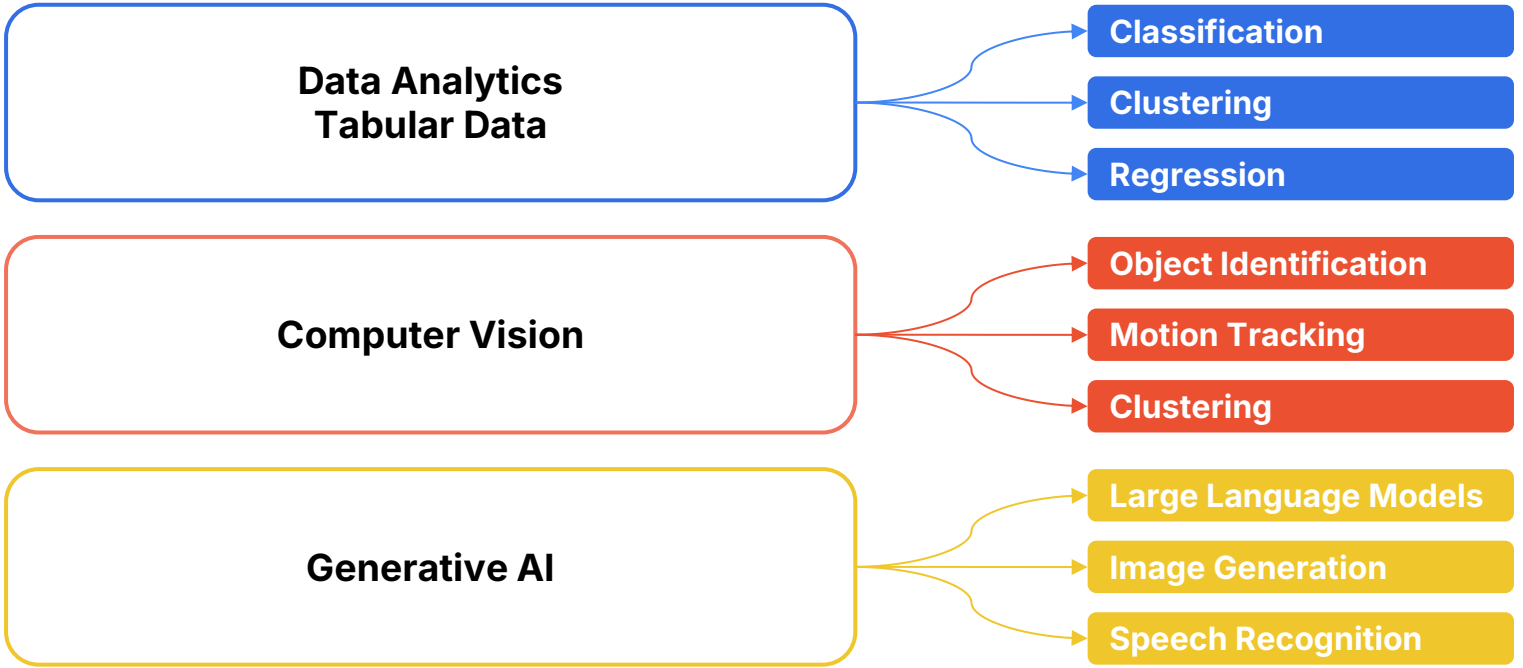
Interplay supports any use case and is not limited to pre-determined processes (e.g. CRM, Case Management) and does not have a self interest on where to run in or which models to use. It helps **enable the solution where it makes most sense** taking privacy, latency and cost into consideration.

It also includes pre-built **"flows"** which are apps. Apps can be modified and customized.

Developers use the studio to build apps quickly. Up to **17x faster** than 'normal' dev-cycles whilst providing **full flexibility for code modification in a language developers know.**



Iterate's Extensive AI Support



Why Iterate?

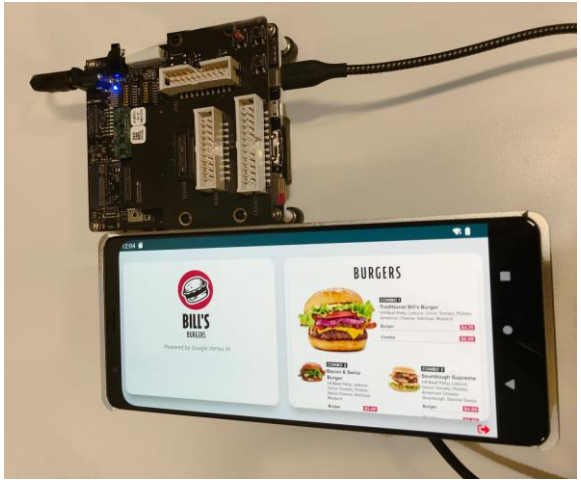
Iterate has unique expertise on small low powered devices and optimizing AI in them and then scaling that via our platform Interplay.

- ➔ **Optimizing vector database on small devices:** Ability to optimize pytorch,non-pytorch dependencies, various search, semantic,keyword, graph algorithms.
- ➔ **Training, prune, quantize models:** Ability to fine-tune, prune and optimize large language models to meet the small device capacity yet meet the needs of the business case.
- ➔ **Ability to optimize inference speeds:** Ability create fast first token generation, batching,speculative and many other optimizations.
- ➔ **One-time R&D scaled by the platform:** Once the R&D is completed the tweaks are applied to Interplay scaling.

We support CPU/GPU and AI PC Configurations

Running Llama3 8B inference in Xeon and training on GPU (A6000)

Running Llama3 8B inference in Core Ultra



Running Llama2/Gemma/Phi on an embedded Qualcomm 8550

Generate

By Iterate.ai

Secure AI Manager Application

What is Generate?

Generate by Iterate.ai is a secure AI Manager application that utilizes RAG and multiple secure vector databases to provide private LLM capabilities to the Enterprise and SMB markets. Generate can run completely privately and securely, either as a locally-installed app directly on an AI PC or via various deployment options within a public or private data center.

Home Consumer

Summarize personal invoices, healthcare records, and more.

SMB Market

Track employee hours, draft contracts, generate invoices

Enterprise Market

Generate RFPs, summarize policies, contract analysis, customer sentiment

Generate Deployment Options

Generate can be deployed through a myriad of options, based on the company's best practices for privacy, data sharing, and self-defined IT best practices.

AI PC Application

- ➔ Deploy app as .exe
- ➔ Completely local on the user's own AI PC
- ➔ Great for Personal or SMB private use, as well as Enterprises upgrading to AI PCs

Local Data Center Appliance

- ➔ Deploy as Docker instance
- ➔ Complies with all data center IT firewalls and security
- ➔ Encrypted database
- ➔ Best for orgs with on-prem data centers

Private Cloud Enterprise Application

- ➔ Deploy as Docker instance
- ➔ Complies with all IT firewalls and security
- ➔ Encrypted
- ➔ Best for remote orgs needing secure private LLM with a private cloud

Public Cloud Secure Instance

- ➔ Deploy as Docker instance
- ➔ Complies with all IT firewalls and security
- ➔ Encrypted
- ➔ Best for remote orgs needing secure private LLM with secure cloud instance

aws

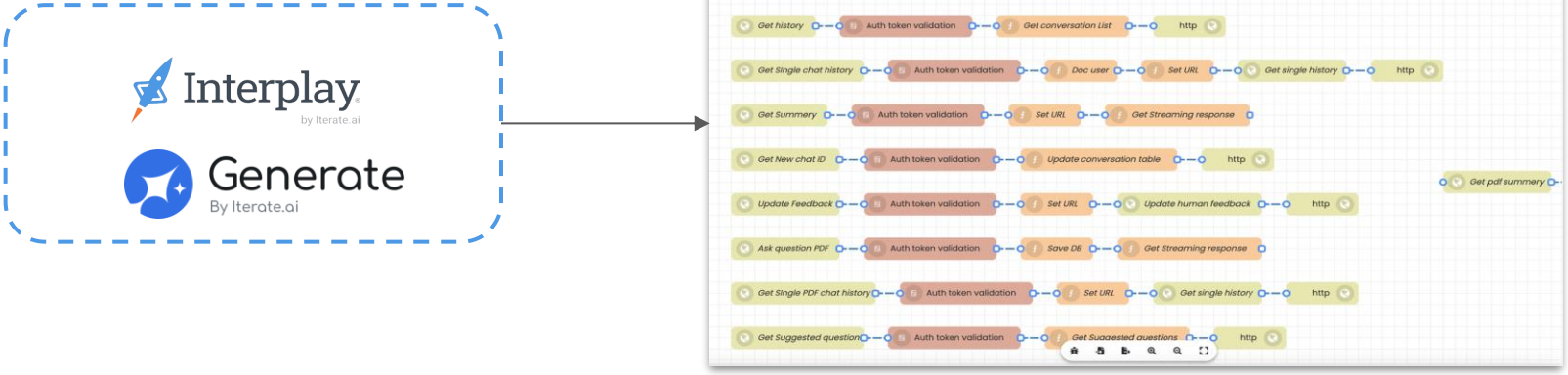
Google Cloud

Azure

Generate Runs on Interplay

As an application built completely on top of Interplay, Generate has the unlimited freedom and flexibility to be customized, improved, and integrated with any use case in mind.

Interplay itself is an industry-tested and proven full-stack AI environment with end-to-end optimizations, from hardware, to code, to compilation, to runtime, to deployment, to scale.



Generate AI PC

By Iterate.ai

Local AI Manager Application

Generate AI PC as a Personal AI Assistant

Generate Personal AI Assistant users can ask questions and get quick answers linked to sources in your private files and documents. Generate efficiently takes advantage of Intel's AI PC architecture to provide the most compelling user experiences and productivity increases.

Generate AI PC is a private LLM manager for the consumer and SMB:

- ➔ Private with no data leaks to outside platforms
- ➔ Configurable to run on different foundation models
- ➔ Completely containerized to run on an AI PC
- ➔ Can run on a private network for highest security
- ➔ Fine-tuned LLM for specific own terminology and personal records
- ➔ Deep document search in conversational interface
- ➔ Integrates into customer service and email queues

SMB Owners

Track employee hours, draft contracts, and more.

Home Consumer

Summarize personal invoices, healthcare records, and more.

Generate AI PC is Secure, Flexible, and Deployable

Secure:

Generate AI PC is built with security at all levels:

- ➔ On-prem deployment
- ➔ locked vector databases
- ➔ encrypted at rest and encrypted access
- ➔ Fine-tuned with prompt jails for business use

Flexible:

Generate AI PC is the most flexible AI management application:

- ➔ Foundation-model agnostic
- ➔ Vendor-agnostic with integrations across all major business suites
- ➔ Large file support
- ➔ RAG and sourced references by default

Deployable:

Generate AI PC is easily deployable across the Intel customer base:

- ➔ Tuned for Meteor Lake and Lunar Lake chips
- ➔ Containerized for on-prem and edge deployable
- ➔ Fits within existing Intel IT relationships with SMB and Personal customers

Generate Demo

Generate AI PC - Dual Roles

Interplay Generate AI PC fulfills a **dual role** for the **Home Consumer** and **SMB** segment.
Generate's flexible RAG engine and local security benefits allow overlapping and unique use cases for both roles.

Ask questions to Health Records

Plan and manage family events

Summarize repair guides and warranties

Conduct personalized learning and education

Summarize and budget expenses

Personal advice and mental wellness

Answer home maintenance and repair tips

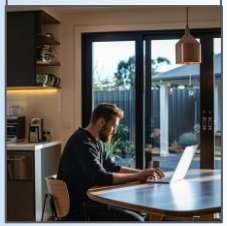
Suggest recipes

Summarize daily news articles

Generate party ideas

Summarize emails and suggest responses

Analyze legal documents

Home Consumer	SMB
	

Summarize invoices

Provide financial advice

Suggest responses to customer queries

Generate and search job descriptions

Generate reports

Summarize meeting notes

Draft contracts and invoices

Summarize financial reports

Analyze budgets and employee hours

Suggest project workflows and timelines

Search product inventory

Analyze payrolls, hours and timesheets

Generate proposals

Generate marketing material

Home Consumer Use Cases | Shared Use Cases | SMB Unique Use Cases

Why AI PC?

Advantage	Benefit
Optimized for Intel	Optimized to Intel Core Ultra for speed. Also deployable end-to-end on other Intel chipsets.
RAG Framework	Removes requirement for costly fine-tunes or GPUs for LLM training. Easy data sourcing, additions, and removals.
Multiple LLMs	Multiple LLMs available. Can deploy and integrate with any open source model.
Completely Local	Does not require internet access. All actions and processing are performed and contained entirely on the device.
Integrations	Integrations with all compatible 3rd party solutions. As an agnostic solution, any solution can be integrated.

Iterate Generate vs Custom Point Solutions

Generate Considerations

- Generate runs on Interplay, a low-code app dev middleware platform with 4000 available integrations
- Generate can run proprietary foundation models with APIs, Open-Source models, or completely local models
- Generate has security incorporated at all levels with encrypted data sources, encrypted vector databases, and encrypted messaging as standard

Point Solutions Considerations

- All-in-one package with limited customization
- Generally limited to pre-packaged integrations
- Proprietary model or ChatGPT wrapper
- Provider must ensure they are up-to-date with evolving LLM landscape and innovation
- Provider may not strictly adhere to security and privacy protocols

Generate avoids vendor Lock-in

Large vendors may encourage or require full integration with their enterprise suite. As the AI arms race continues and acquisitions and investments continue to be made, it is difficult to discern which vendor is the 'best' overall.

Private/Custom Providers	Microsoft	Google	amazon
Cloud Suite	Azure	Google Cloud	aws
LLM	ChatGPT	Gemini	Amazon Bedrock

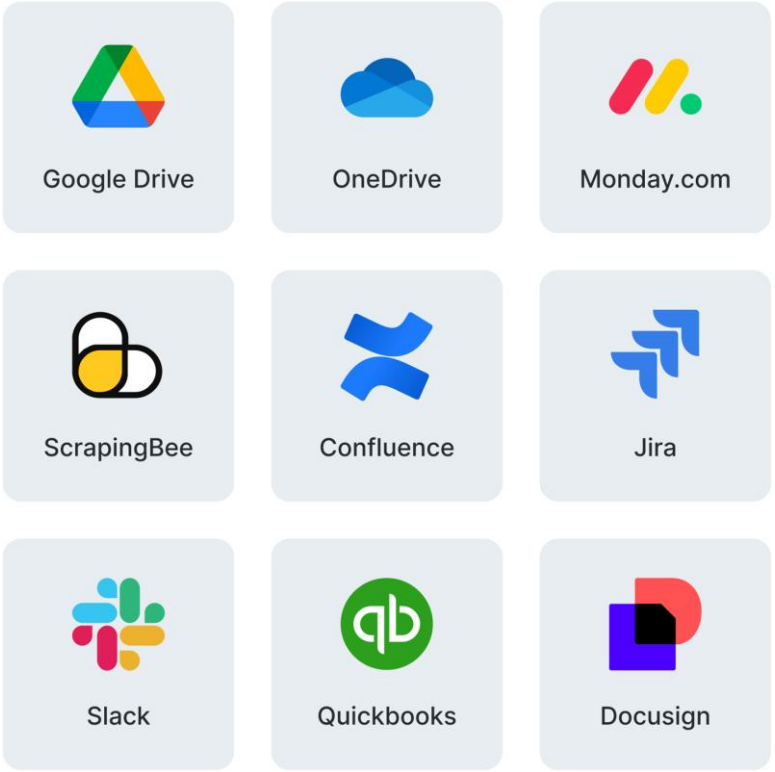
Generate Application - OEM Deployment

Generate AI PC is both powerful and highly efficient. That's why Intel chose Generate for AI PCs.

On millions of AI PCs powered by Intel chipsets, Generate will ship on the toolbar.

- **Local RAG:** it will index thousands of photos and documents.
- **Private:** It can keep private information private.
- **LLM options:** You can use a variety of LLMs, from Meta's LLaMA2 to Phi to Mistal to GPT-4.
- **Business and Personal:** It can satisfy the needs for both business and personal users.

Integrations



Interplay Generate AI PC connects and pulls in data from any 3rd party solution that has an API or other general means of access.

Generate AI PC is platform-agnostic. The application is not directly subsidized or tied directly to any single vendor or platform. Iterate.ai can integrate any 3rd party solution into the Generate application.

Current integrations include:

- Google Drive suite (Docs, Slides, Sheets, Gmail)
- Onedrive suite (Docx, Powerpoint, Excel)
- Common workflow managers (Monday, Slack)
- Common Ticket trackers (Jira, Confluence)

Generate Private RAG

Supported Data Types

Local Upload

Generate

AI PC

By Iterate.ai

Local RAG and LLM

Generate AI PC Features

- LLM General Q&A
- LLM Direct Source Q&A
- LLM Source Summaries
- LLM Transformations
- Image Search

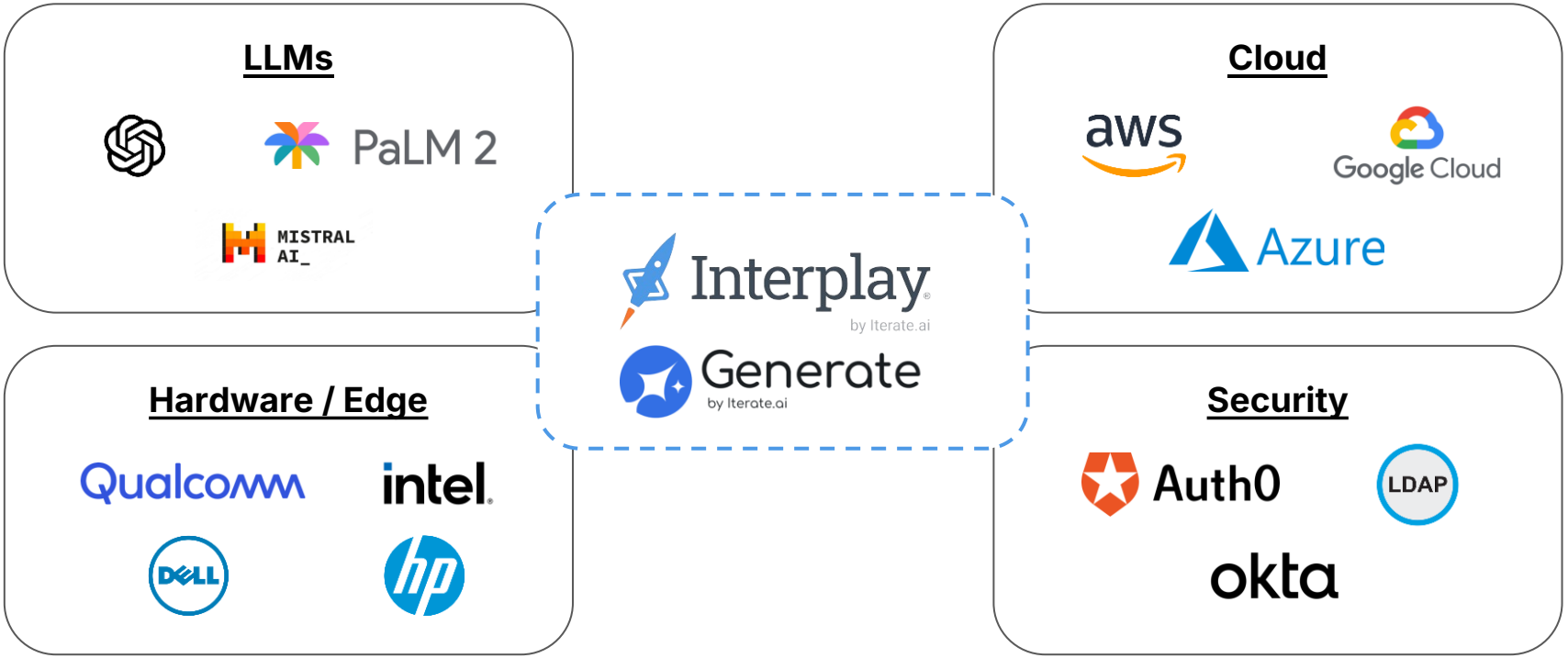
Generate Enterprise

By Iterate.ai

Local, Private, and Secure LLMs for Organizations

Generate avoids vendor Lock-in

Generate Enterprise is model-agnostic and platform-agnostic, meaning that it can run many different AI foundation models on many different platforms and environments. This assures that Generate can always bring best-of-breed in the rapidly evolving AI market.



Generate Enterprise Use Cases

	Executive Teams	Marketing Teams	Legal Teams	Research Teams	Operation Teams	HR Teams
Use Case	● Summarize reports ● Automate email workflows ● Extract insights through LLM Q&A	● Summarize reports ● Refine and retone complex documents ● Craft branded marketing copy ● Automate marketing workflows ● Conduct data-driven campaign strategies ● Retrieve salient data points and statistics ● Extract trends/insights through LLM Q&A	● Summarize reports ● Refine and retone complex documents ● Combine company context data ● Retrieve salient data points and statistics ● Extract insights and strategies through LLM Q&A with company data and legal concepts	● Summarize complex research reports ● Refine and rephrase documents ● Quickly retrieve salient data points and statistics ● Extract insights through LLM Q&A ● Build research curations with LLM analysis	● Automate customer service workflows ● Summarize operational guides and manuals ● Extract insights with customer feedback and service metrics through LLM Q&A	● Quickly retrieve and summarize company policies through LLM Q&A ● Refine and retone job descriptions ● Automate internal email communications ● Extract insights among workplace data to inform future strategy through LLM Q&A
Document Types	● Board meeting summaries ● Corporate communications ● Pitch and sales decks ● Sales forecasts ● Revenue projections ● Market analysis reports ● Financial statements ● Industry whitepapers	● Market research reports ● Customer surveys ● Competitor analysis ● Campaign performance metrics ● Ad copy ● Social media posts ● Email newsletters ● Website content	● Case law precedents ● Statutes and regulations ● Legal briefs ● Contracts and agreements ● Contract clauses	● Academic journals ● Research articles ● Conference papers ● Theses and dissertations ● Data repositories ● Data analysis reports	● Standard operating procedures (SOPs) ● Training manuals ● Incident reports ● Inventory lists ● Training materials ● Internal memos ● Operational guidelines ● FAQ documents ● Troubleshooting guides ● Equipment manuals	● Employee handbooks ● HR policies and procedures ● Employment contracts ● Benefits summaries ● Job descriptions ● HR communications ● Leave request forms ● Workplace safety guidelines ● Diversity and inclusion policies

Technical Specifications

IPEX → OpenVINO Optimizations

Through collaboration with the OpenVino team, Iterate switched from IPEX to OpenVino for its runtime environment, and achieved significant improvements in RAM usage and device flexibility.

	IPEX-LLM	→ OpenVINO™
RAM Usage 8B Model	20-22GB	8-9GB
RAM Usage 1.5B Model	14-16GB	4-6GB
Application Size	10GB	5GB
Compatible Devices	Core Ultra H-sku	Core Ultra U-sku and H-sku All OpenVINO supported Intel devices

Other optimizations

- ➔ **LLM On / Off Toggle:** Generate allows users to turn the LLM On and Off from their device, seeing up to a 3x improvement in GPU wattage.

- ➔ **Fine-tuned SLMs:** Generate offers Interplay RAG Pro and Interplay RAG, a custom fine tuned 8B and 1.5B model with increased analysis and reasoning capabilities.

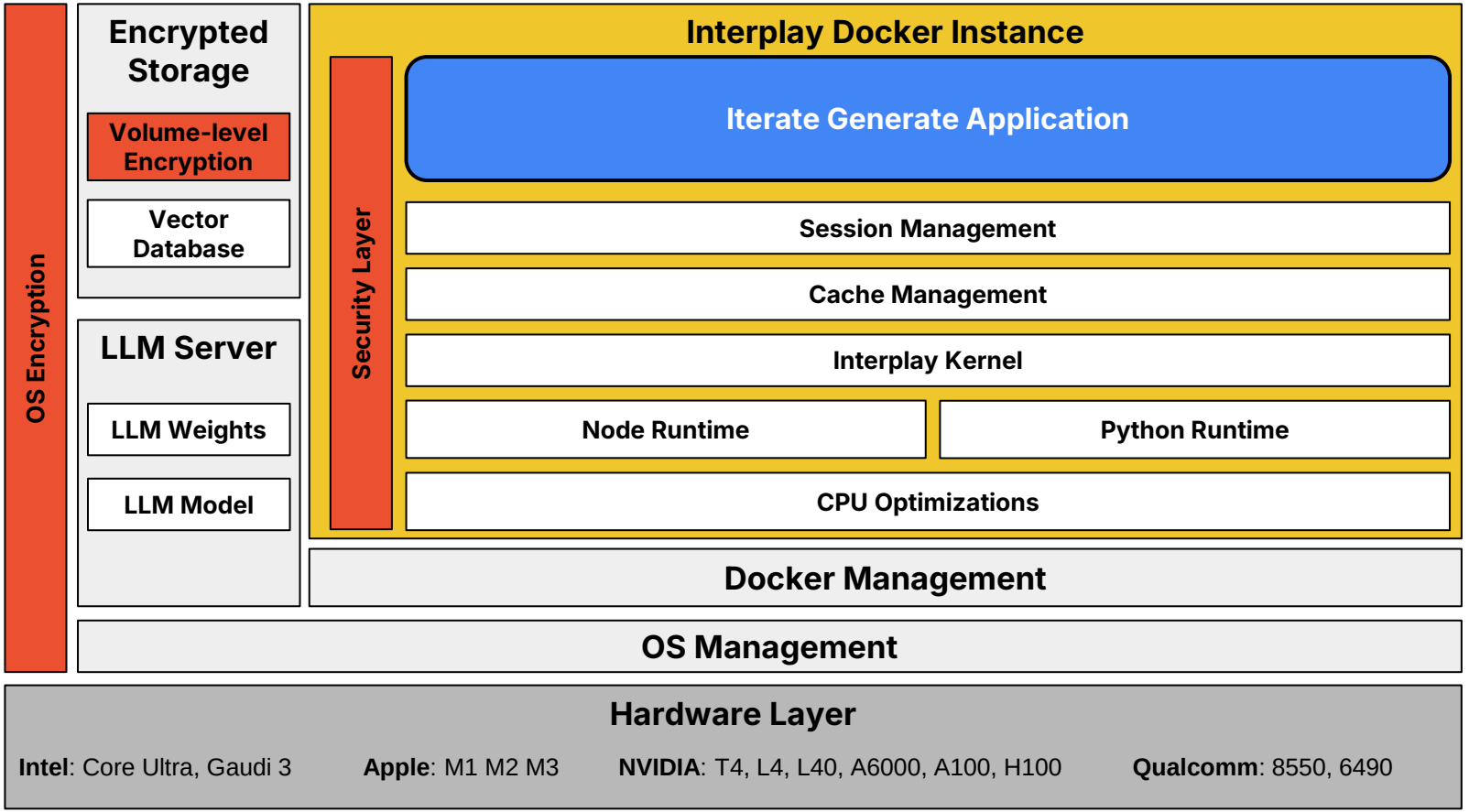
- ➔ **Lunar Lake:** Generate is compatible with the newest Lunar Lake platform, seeing up to a 3x improvement in first token latency, and 2x improvement in TPS.

Benchmarks

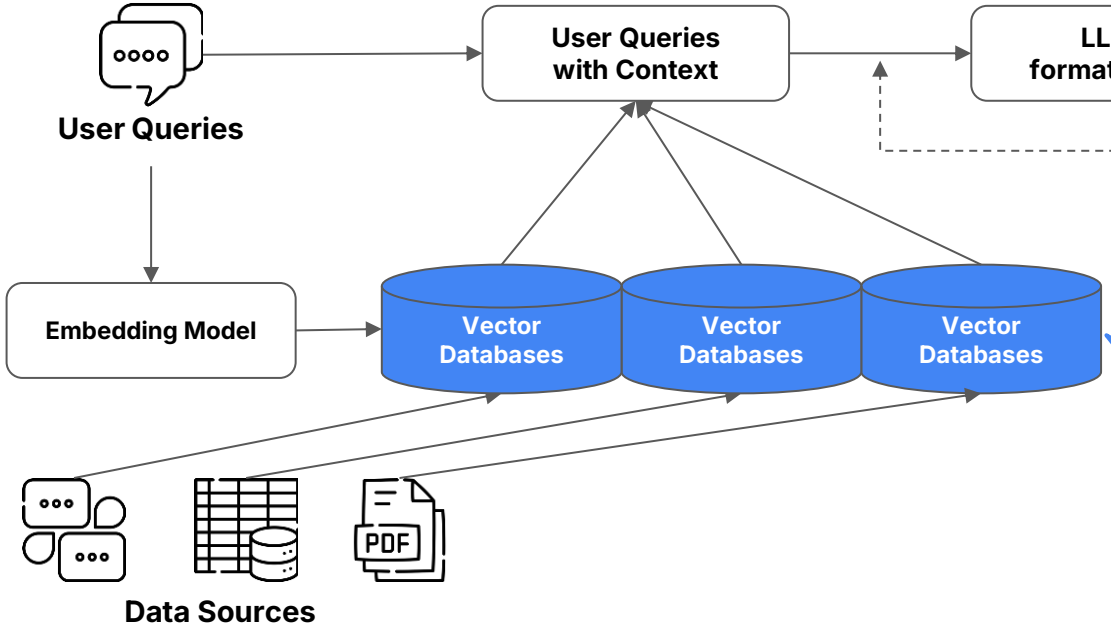
Average Tokens Per Second

*Tested June 2024 using Llama8B and Qwen1.5B SLM workload running 1024-token document summarization loops on default prompt settings on the Core Ultra AI PC 32GB RAM and Intel 155H CPU (integrated CPU, GPU, NPU) configuration.

Generate Architectural Stack



Iterate Generate utilizes RAG with separate vector databases



With separate encrypted databases for each data source, Generate improves RAG accuracy, reduces hallucinations, and allows role-based segmented answers for each user in the organization.

Technical Differences between Generate AI PC and Enterprise

Generate AI PC By Iterate.ai		Generate Enterprise By Iterate.ai
Home Consumer and SMB	**Audience**	Enterprise Teams
Document Search, Image Search	**Available Applications**	Document Search, Email Service Pilot, Text Augmentation
3rd Party Integration Support, Completely local and offline LLM	**Features**	Security Roles, User Access, Document Restrictions, Monitoring, Advanced OCR
Dall-E Generation (Cloud), Interplay RAG Pro(Local LLM), Interplay RAG (Local LLM), BLIP (Local LLM)	**LLMs**	Public Cloud: GPT4o / 4 / 3.5 Private Cloud: Llama8B, Mistral7B, Gemma2B
Local Executable	**Deployment**	Private Cloud Deployment, Local Docker deployment, Enterprise Data Centers
150mb (without LLM)	**Footprint**	60gb

Questions & Answers

Thank you

Dream **big**, build **fast**.

Iterate.ai
Dream **big,** build **fast.**

Interplay
By Iterate.ai