TECHNICAL WHITE PAPER: September 2025

Unlocking the Power of GenAI

with VMware Private AI Foundation with NVIDIA and DKube

Table of contents

Private GenAI: Introduction

Generative AI (GenAI) represents a paradigm shift in enterprise computing. By leveraging large-scale foundation models trained on vast datasets, organizations can automate complex tasks, generate new content, and derive strategic insights from data previously considered too unstructured or complex.

However, with great power comes great responsibility. Deploying GenAI at scale brings questions of data security, compliance, ethical usage, model reliability, and return on investment. This whitepaper aims to demystify the process of building GenAI applications within enterprise boundaries by highlighting the combined strengths of VMware Private AI Foundation with NVIDIA and DKube's AI services.

Together, Broadcom and DKube enable organizations to navigate technical, operational, and strategic barriers, accelerating the journey from AI infrastructure investment to real-world business outcomes. We'll explore how to select the right foundation model, prepare and secure data, enforce application-level safeguards, evaluate AI performance, and deploy confidently in private environments.

Private GenAI: The Challenge

Using cloud-based GenAI APIs seems simple: plug in your data and receive outputs. But when data privacy, regulatory compliance, and organizational security become priorities, things grow complex quickly. Enterprises must think beyond proofs-of-concept toward scalable, secure, and compliant production deployments.

The challenge is threefold:

1. Data Governance: Enterprise data often includes personally identifiable information (PII), trade secrets, or regulated content. Sending this to external APIs may violate compliance policies (For example, GDPR, HIPAA).
2. Lack of AI Talent: Organizations often lack professionals who understand model fine-tuning, prompt engineering, observability, and deployment automation.
3. Infrastructure-Outcome Gap: After investing in GPUs, storage, and software licenses, many enterprises struggle to realize tangible AI value due to integration complexity and lack of application development expertise.

This is where Broadcom and DKube provide great value to customers. Broadcom provides the secure private cloud infrastructure and DKube offers expert teams to build, deploy, and optimize applications.

Building Private AI Applications

1. Selecting the Right Foundation Models (FM)

FMs are deep neural networks trained on large datasets to perform tasks across natural language, vision, audio, and multimodal inputs. Choosing the right FM is the first—and arguably most critical—step in GenAI development. It may become an iterative experimentation process in many cases but following guidelines before starting may reduce the effort and time required.

- Use Case Alignment: The model must support the required use case: summarization, search, question answering, document generation, etc.
- Performance Benchmarks: There are many public evaluations of models available to compare accuracy, coherence, latency, and safety across models like GPT, LlaMA, Mistral, and Claude.
- Open vs. Closed Models: Open-source models (e.g., LlaMA, Mistral) offer transparency and customization but require more ops support. Closed models (e.g., GPT) may be more accurate but have restrictive licensing.
- Token Handling & Context Length: Applications involving long documents or user history require models with high context windows (e.g., 32k+ tokens).
- Model Licensing & Cost: An open-source model may be free to use but every query still consumes GPU and licensed software resources and hence cost. A model has to be efficient in the compute usage for the task involved.

Choosing a model optimization strategy is just as critical as selecting the model itself, especially for latency-sensitive applications such as chatbots or virtual assistants.

- Quantization: Reduce model size and improve inference speed by using lower precision (e.g., INT8 instead of FP32).
- Distillation: Create a smaller model (student) from a larger one (teacher) while preserving performance.
- Model Sharding and Parallelism: Split large models across multiple GPUs for faster performance in production.

2. Deployment Considerations

Fine-tuning open-source models (e.g., LlaMA, Mistral) allows organizations to embed proprietary knowledge directly into the model without external API calls. However, it may not be required with every application in an organization. There are simpler ways to use the models before fine tuning is attempted.

- Zero-shot Learning: The model is prompted to perform tasks without additional training. This works well with strong general-purpose models.
- Few-shot Learning: The model is given a few examples in the prompt to improve performance.
- Fine-tuning: The model is retrained on a targeted dataset to deeply embed domain knowledge (e.g., legal documents, insurance claims). This can drastically improve task accuracy but requires a process to update the model.

Retrieval-Augmented Generation (RAG)

RAG enhances a generative model by injecting external knowledge into the generation process. It combines retrieval-based and generative approaches to produce more accurate, factual, and context-aware responses.

Key Components of RAG

- Retriever: Searches a knowledge base (e.g., documents, embeddings) for relevant content
- Generator: A language model that uses retrieved content to generate the answer
- Knowledge Store: Often a vector database or search engine (e.g., FAISS, Elasticsearch)

How RAG Works (Simplified Flow)

- User asks a question (e.g., “What are the symptoms of diabetes?”)
- The retriever finds relevant documents from a corpus using vector similarity search
- The generator receives the original question plus retrieved documents as input
- It generates a factually grounded answer based on both

Benefits of RAG

- Grounded responses — relies on up-to-date, verifiable sources
- Domain adaptation — useful in specialized fields (e.g., legal, medical)
- Smaller models can compete — thanks to external knowledge injection
- Explainability — you can trace outputs back to source documents

Applications of RAG

- Enterprise QA systems
- Legal/medical assistants
- Customer support
- Scientific research tools
- Internal document search and summarization

Example

Without RAG: GPT might hallucinate an answer about a recent event it wasn't trained on.

With RAG: It pulls documents about the event from an internal news corpus, then summarizes them accurately.

Once a foundation model is selected, enterprises must decide whether to use it as-is (zero-shot or few-shot) or to fine-tune it with domain-specific data.

3. AI Agents for Data Preparation

High-quality data is the lifeblood of successful GenAI applications. Poor data leads to poor outputs, no matter how powerful the underlying model. AI agents can drastically improve data readiness by automating repetitive, error-prone steps in the pipeline.

Key Agent Types

- Data Annotation Agents: These agents label text, images, or structured data with metadata, classification tags, or context-specific descriptors.
- Data Validation Agents: These agents scan datasets for missing values, inconsistencies, or misclassifications.
- Data Transformation Agents: These agents normalize, tokenize, or reformat data to match input specifications.
- Data Augmentation Agents: These agents generate synthetically enhanced samples to reduce model bias or expand underrepresented categories.

Benefits of AI-Driven Preparation

- Efficiency: Reduces manual effort by 60–80%.
- Data Quality: Identifies and corrects errors early.
- Customization: Enables contextual preprocessing based on task.
- Scalability: Supports large-scale pipelines across domains.

Best practices include maintaining data lineage, using human-in-the-loop checks, and testing impact of data changes on model outputs.

4. Model Accuracy and Evaluation

Once a model or a set of models are initially selected, deployment considerations are finalized such as RAG or no RAG (zero shot, few shot training) and the data prep agents are built, the final outputs have to be evaluated during the early experimentation process. The accuracy of large language models (LLMs) can be defined and measured in different ways depending on the task, domain, and evaluation framework.

Task Type Common Accuracy Metric

Accuracy metrics vary by the type of task the AI application is supposed to perform.

Text classification	Exact match accuracy, F1 score
Multiple-choice QA	% of correct answers
Open-ended QA	Exact Match (EM), F1 score
Summarization	ROUGE, BLEU, METEOR
Code generation	Pass@k, exact match
Dialog/Chat	Human eval, dialog coherence, helpfulness
Translation	BLEU, chrF++

Here’s a breakdown of some of the metrics listed above:

F1 score: It is the harmonic mean of precision and recall:

- **Precision** = Correct positive predictions / Total predicted positives
  - “Of all the results the model said were positive, how many actually were?”
- **Recall** = Correct positive predictions / Total actual positives
  - “Of all the truly positive items, how many did the model catch?”
- **F1 Score** = 2 x (Precision x Recall) / (Precision+Recall)

BLEU (Bilingual Evaluation Understudy): It is an automatic metric for evaluating the quality of machine-generated text, especially in tasks like machine translation, text summarization, and paraphrasing. BLEU evaluates how similar a machine-generated sentence is to one or more human-written reference sentences, based on:

- **n-gram overlap** (e.g., word pairs, triples)
- **brevity penalty** (to discourage overly short outputs)

How BLEU Works:

1. Tokenize both the generated and reference text.
2. Count overlapping n-grams between the generated text and the references.
3. Compute precision for unigrams (1-grams), bigrams (2-grams), etc.
4. Apply a brevity penalty if the generated sentence is shorter than the reference.
5. Combine all using a geometric mean
6. Score Range: 0 to 1 (or 0% to 100%). Higher is better.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): It is a set of metrics used to evaluate automatically generated text, particularly for summarization and paraphrasing.

What It Measures:
ROUGE focuses on recall—how much of the reference text is recovered in the output.

Common Variants:

- ROUGE-N: Overlap of n-grams (ROUGE-1: unigrams, ROUGE-2: bigrams)
- ROUGE-L: Longest Common Subsequence
- ROUGE-W: Weighted LCS (penalizes non-consecutive matches)
- ROUGE-S: Skip-bigram (non-adjacent word pairs)

ROUGE-N Formula:
ROUGE-N = (Number of overlapping n-grams) / (Number of n-grams in reference)

ROUGE vs BLEU:

- BLEU emphasizes precision; ROUGE emphasizes recall.
- BLEU is popular for translation; ROUGE is widely used in summarization.

METEOR (Metric for Evaluation of Translation with Explicit Ordering): It is an automatic evaluation metric for machine-generated text, originally designed for machine translation but also applicable to summarization and text generation.

What It Measures:
METEOR evaluates how closely a generated sentence matches one or more reference sentences. It combines both precision and recall but gives more weight to recall.

Key Features:
- Uses flexible matching: exact, stemming, synonyms, and paraphrases
- Prioritizes recall with a weighted F-score:
- F_mean = (10 * Precision * Recall) / (Recall + 9 * Precision)
- Applies a fragmentation penalty to discourage disordered matches

Score Range:
0 to 1 (or 0% to 100%). Higher scores indicate better semantic and syntactic alignment with the reference.

Comparison with BLEU and ROUGE:
- BLEU focuses on precision and n-gram overlap
- ROUGE emphasizes recall and n-gram or subsequence matches
- METEOR balances precision and recall, and includes semantic matching via stemming and synonyms

Example:
Reference: “A quick brown fox jumps over the lazy dog.”

Generated: “The fast brown fox leapt over a sleepy dog.”

METEOR scores this relatively high due to synonym and stem matches, even if exact words differ.

LLM Accuracy in Complex or Open-Ended Tasks
For more nuanced tasks (e.g., reasoning, summarization, multi-turn dialogue), measuring accuracy with the metrics above may not be enough. More testing is required.

- **Factual Consistency:** Is the output factually correct?
- **Faithfulness:** Does the output accurately reflect the input?
- **Completeness:** Are all relevant parts of the task addressed?
- **Alignment with instructions:** Does the output follow task-specific instructions?

LLMs are often evaluated using **human raters** or **proxy tasks** that reflect real-world usage.

Practical Accuracy Considerations
- **Prompt sensitivity:** LLMs may perform better or worse depending on how you ask.
- **Hallucinations:** Factually inaccurate content presented confidently.
- **Model size and training data:** Typically correlate with accuracy—but not always linearly
- **Toxicity:** Frequency of unsafe or biased outputs

Enterprises may also fine-tune models for their specific domain to improve performance while preserving general language understanding.

To properly evaluate a private GenAI application, organizations must create representative evaluation datasets:

Evaluation Dataset Creation

Customized test sets should be created with SME (subject-matter expert) input to simulate real-world tasks.

- **Legal:** Accuracy of citations, correct use of legal precedent, adherence to formatting standards.
- **Healthcare:** Compliance with clinical guidelines, absence of hallucinated diagnoses or medications.
- **Customer Support:** Resolution effectiveness, empathy/tone, alignment with company policy.

General metrics (BLEU, ROUGE, F1) may not reflect real-world utility in specific domains. Consider customizing evaluation for context:

Domain-Specific Evaluation

Even high-performing models need validation. Evaluating both individual components and end-to-end AI systems ensures they deliver meaningful results.

Foundation Model Evaluation

- Benchmark across known datasets.
- Test for bias using ToxiGen or similar tools.
- Compare against business KPIs (e.g., support resolution rate, document summarization time).
- Human Ratings, Latency, Task Success Rate, Robustness Tests (adversarial inputs)

Securing and Monitoring AI Applications in Production

Security is a top concern when deploying GenAI in production, especially when handling sensitive or proprietary data. Unlike the development phase the production phase could last years and hence the monitoring for AI security vulnerabilities and quality of AI generated content/answers becomes very important. It requires a different set of techniques than what was used in development.

Key Security Components

- **Data Encryption:** Encrypt data at rest and in transit using standards like AES-256 and TLS 1.3.
- **Access Controls:** Implement granular RBAC and audit logs to prevent misuse.
- **Model Integrity:** Use checksums, version control, and signing to prevent model tampering.
- **Runtime Monitoring:** Log inference activity, flag anomalies, and trace decisions.
- **Policy Compliance:** Ensure alignment with internal and external governance (GDPR, HIPAA).

Example Threats During Long Term Production Use

- **Prompt Injection:** Malicious user input alters model behavior.
- **Model Poisoning:** Training data is manipulated to produce biased or harmful outputs.
- **Data Leakage:** Outputs inadvertently expose sensitive training data.

This requires using monitoring tools like Langfuse, Loki, and Prometheus to track and debug inference sessions in real time.

Model and Application Accuracy and Evaluation

Frameworks like Holistic Evaluation of Language Models (HELM) provide templates for such evaluations.

- **Fairness:** Ensure outputs do not disproportionately impact protected groups.
- **Explainability:** Gauge how understandable outputs are to non-technical users.
- **Safety:** Detect outputs that may be offensive, toxic, or illegal.

Evaluation should go beyond just accuracy and should be automated and integrated into the CI/CD process, triggering new evaluations when data or AI model is revised

Integration of Evaluation in CI/CD

This ensures performance consistency and early detection of regressions.

- A new model version is deployed.
- Major data updates are made.
- Application logic is changed.

A dataset of 100–200 queries per use case is a good starting point, with manual ratings from SMEs to establish a ground truth.

- **Synthetic Queries:** Create user-like queries using LLMs and validate them.
- **Historical Logs:** Anonymize and repurpose past support tickets or emails.
- **Edge Cases:** Include adversarial, ambiguous, or multi-intent examples.

Custom/Advanced Techniques

Depending on the application domain following techniques can be used:

- **Rubric-Based Evaluation:** Define qualitative axes (e.g., Helpful, Harmful, Complete) and score outputs accordingly.
- **Automated and Human Evaluation:** LLM-as-a-Judge vs. Human-in-the-Loop
  - **LLM-as-a-Judge:** This involves the use of LLMs themselves as evaluators—a paradigm referred to as LLM-as-a-judge. In this approach, the language model assesses its own or other AI-generated outputs based on predefined criteria such as accuracy, coherence, or relevance. These evaluations may involve direct comparisons with ground-truth data or rely on statistical measures such as perplexity and F1 score listed above. Also, one should consider using an LLM to grade the generated outputs based on prompts with scoring rubrics.
  - In contrast, **human-in-the-loop** evaluation incorporates human judgment into the assessment process. This method is particularly valuable when evaluating subjective attributes such as fluency, contextual appropriateness, and overall user experience—dimensions that automated metrics may struggle to quantify effectively. Generally starting with 30-40 questions from human experts and how the AI application answers can give a good indicator on the quality of the AI application.

A combination of the two evaluation techniques can lead to better-performing private AI applications.

Accelerating GenAI with DKube

As one can see above building private Gen AI applications that provide high accuracy and align with the values and goals of an organization can be a highly specialized skill in addition to the base foundation of compute, GPU, virtualization, Kubernetes and other hardware/software infrastructure. Many enterprises do not have people with these skills. This is where DKube team can help.

DKube, an AI services company headquartered in San Jose, CA, has been building enterprise AI applications for clients across multiple sectors for many years. DKube brings its skilled AI team to clients so they can rapidly build and deploy Private AI applications for enterprises with VMware Private AI Foundation with NVIDIA. The delivery model is tailored for fast ROI from AI investments. The focus is on speed, business outcomes, and resource efficiency -compute, storage, people. DKube team has been porting several of its AI applications to run on VMware Private AI Foundation with NVIDIA. Examples and use cases include industries such as legal, health insurance, life sciences, mortgage banking applications. These AI applications enhance functions such as customer support, search engines, document processing, pharma drug discovery, research assistants, business intelligence etc.

DKube offers a comprehensive AI services team that partners with clients to deliver GenAI applications in 12 weeks. The model includes:

- **Staffing:** Prompt engineers, data engineers, performance optimizers, solution architects
- **Base Stack:** Open-source foundation (Langfuse, OpenWebUI, Grafana) tailored to client needs
- **Deployment:** Integrated with existing authentication, RBAC, and security systems for a persona based access for development and production use.
- **Customization:** Each stack is configured per department or use case, and clients retain all code and IP

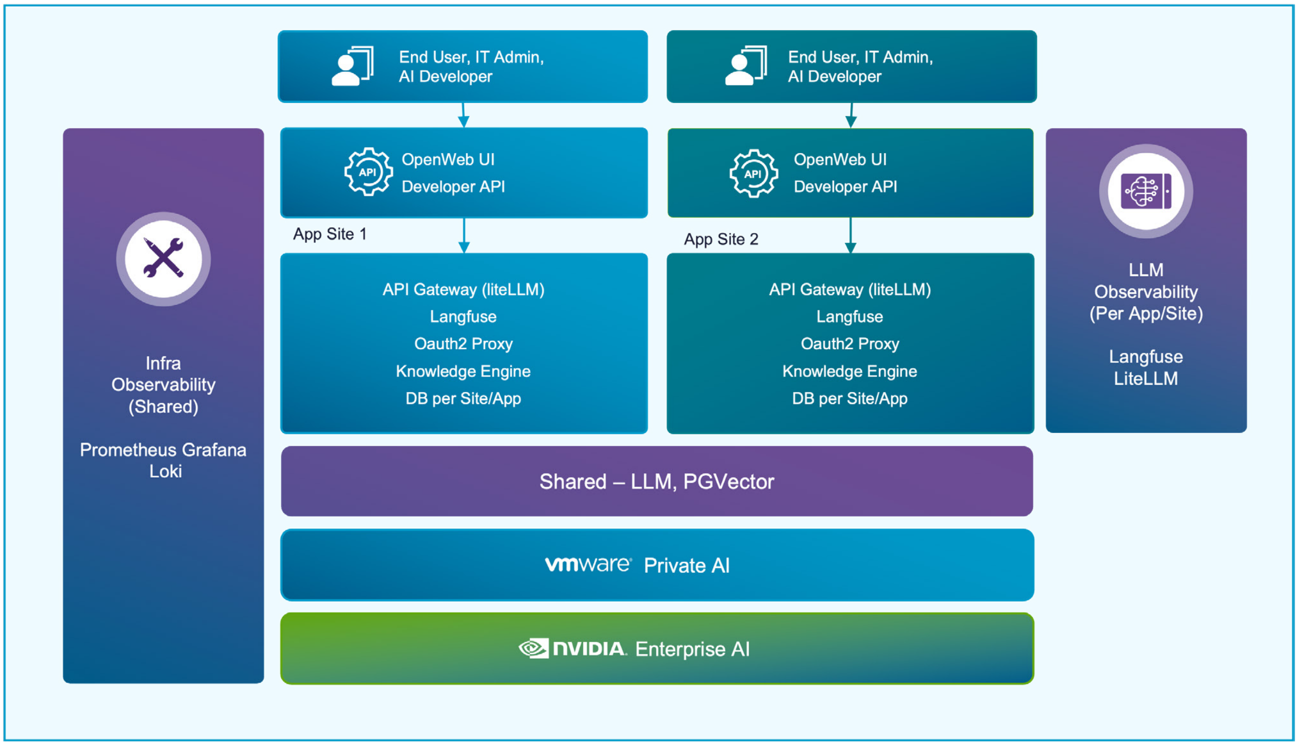

Figure 1: The different technical stacks and customer personas that come together between Nvidia, VMware and DKube to deliver on this promise.

The base stack is integrated with the client's authentication at dept and/or site level as necessary within a week or two. Depending on the client not all components may be needed whereby DKube custom fits the right components for each client. This dramatically improves the speed of application development for the client with their own data. The base stack is offered under a fixed term service contract. The client pays for no additional software license(s) and keeps all the code and configuration files developed at the client side.

In addition to the base stack DKube team brings a well-oiled and repeatable process with milestones for AI application development and evaluation as listed in this blog above.

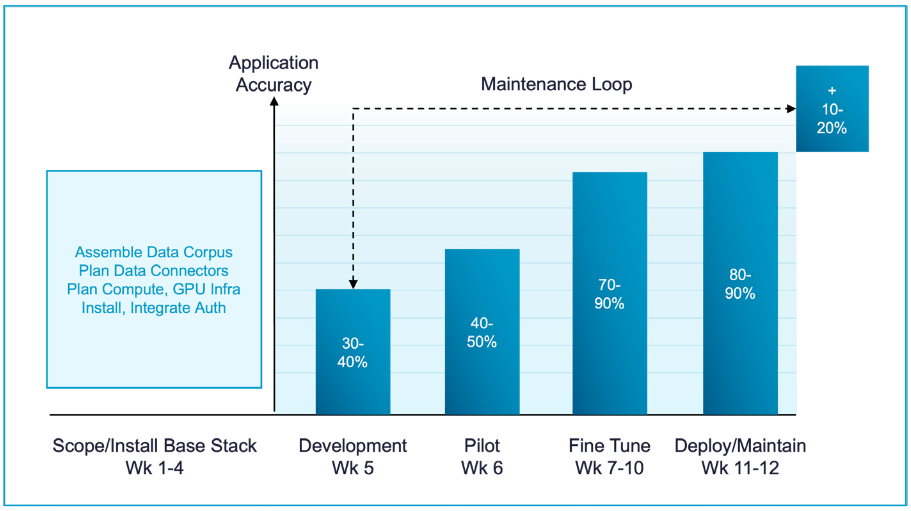

Figure 2: shows more details on this 12-week process.

Conclusion

Broadcom and DKube empower enterprises to unlock GenAI's potential without sacrificing privacy, control, compliance or agility. VMware Private AI Foundation with NVIDIA ensures secure model hosting and data governance. DKube's AI services team accelerates time-to-value by bringing in application expertise. To make the customer experience as streamlined and efficient as possible, DKube and Broadcom have partnered together to help customers build on the solution. Dkube's teams have been validating and optimizing workloads on VMware Private AI Foundation to provide the confidence and agility customers demand for their Enterprise AI workloads.

Through careful model selection, robust data preparation, accuracy and evaluation metrics, security enforcement, and continuous evaluation, enterprises can deliver AI applications that are not only powerful—but also responsible, secure, and aligned with business goals.

Ready to get started with VMware Private AI Foundation with NVIDIA and DKube?

- Complete this form to contact us!
- Read the VMware Private AI Foundation with NVIDIA solution brief.
- Learn more about VMware Private AI Foundation with NVIDIA.

Authors

Alex Fanous,
VMWare by Broadcom

Alex Fanous is a seasoned professional on the VCF Advanced Services Architecture Team at VMware by Broadcom. With extensive experience in cloud infrastructure and virtualization, Alex focuses on designing, implementing, and optimizing advanced solutions for VMware Cloud Foundation (VCF). He collaborates with enterprises to drive innovation and efficiency, leveraging cutting-edge technologies to meet complex business needs. Alex's expertise ensures seamless integration and robust performance across enterprise environments. Passionate about shaping the future of running AI apps on-prem and driving cutting-edge solutions at VMware by Broadcom's Advanced Architecture Team. He is focused on our VMware Private AI Foundation with NVIDIA solution.

Ajay Tyagi,
Dkube.IO

Ajay Tyagi leads the market and business development at DKube.IO, an enterprise AI productivity solutions and services company based in San Jose, CA. Ajay has been the founding member of DKube and has deep understanding of the AI applications development process. Prior to DKube, Ajay built or promoted server hardware and software product lines at names like Intel, Dell, Broadcom for different vertical markets esp telcos and banking.

The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries. For more information, go to www.broadcom.com. All trademarks, trade names, service marks, and logos referenced herein belong to their respective companies. Broadcom reserves the right to make changes without further notice to any products or data herein to improve reliability, function, or design. Information furnished by Broadcom is believed to be accurate and reliable. However, Broadcom does not assume any liability arising out of the application or use of this information, nor the application or use of any product or circuit described herein, neither does it convey any license under its patent rights nor the rights of others.

Item No: bc-vmw-template-whitepaper-tech-A4-v1 9/25