Enterprise AI: LLMs, RAG, and Agentic AI Innovation with Governance

Executive Summary

Large Language Models (LLMs) have become the foundation of modern AI, transforming industries with their ability to generate human-like text and support intelligent automation. They have unlocked new opportunities in customer engagement, operations, and decision-making, but enterprises increasingly require solutions that extend beyond static knowledge to deliver adaptability and domain-specific intelligence.

Retrieval-Augmented Generation (RAG) addresses this need by combining LLMs with external, dynamic knowledge sources. This approach enhances accuracy, relevance, and scalability while reducing reliance on retraining. The next step in this journey is Agentic AI, which moves beyond retrieval to incorporate autonomous reasoning, planning, and tool use. These systems strengthen adaptability while embedding compliance and oversight into execution. To realize their full potential, enterprises must adopt governance-by-design frameworks that ensure trust and accountability.

This paper discusses Generative AI and the CMPak AI+ Strategy, exploring the evolution from Large Language Models (LLMs) to enterprise-grade Retrieval-Augmented Generation (RAG) systems, the progression toward Agentic AI, and the governance framework that ensures trustworthy, scalable, and compliant AI adoption.

1. AI Evolution in Practice: LLMs, RAG, Agentic AI, and Governance

Large Language Models (LLMs) have emerged as the foundation of modern AI, transforming how organizations harness information and deliver intelligent services. Their ability to understand and generate coherent, context-aware language has unlocked diverse applications, from conversational agents and digital assistants to automated content generation and code development. LLM adoption is accelerating globally as enterprises leverage these models to streamline operations, enhance customer engagement, and create new opportunities for innovation. By enabling natural, human-like interactions, LLMs are redefining productivity and shaping the future of digital business.

Retrieval-Augmented Generation (RAG) builds on this foundation by combining LLMs with dynamic external knowledge sources. Deployed worldwide in enterprise knowledge assistants, and policy-driven chatbots, RAG grounds responses in up-to-date information, reducing errors and improving trust. Its capacity to scale across multiple knowledge domains makes it particularly valuable in sectors such as healthcare, telecom, finance, and law, where accuracy and compliance are critical. By eliminating the need for frequent fine-tuning, RAG enhances adaptability while lowering costs, positioning it as a practical solution for enterprises operating in fast-changing environments.

Agentic AI represents the next stage of this evolution, extending beyond retrieval to enable systems with reasoning, planning, and autonomous tool use. These agent-driven architectures are being applied globally in areas such as procurement automation, compliance monitoring, workflow orchestration, and research acceleration. Agentic AI can break complex problems, select optimal retrieval strategies, and integrate external tools or APIs, delivering richer and more context-sensitive outcomes. As enterprises pursue greater automation and resilience, Agentic AI is emerging as a transformative force driving efficiency, performing complex actions autonomously, and empowering strategic decision-making.

In parallel, governance has become the cornerstone of responsible AI adoption. Across industries, organizations and regulators are converging on frameworks that prioritize transparency, accountability, and regulatory compliance. Governance-by-design embeds these safeguards directly into the development and deployment lifecycle, ensuring that AI systems are not only innovative but also reliable and trustworthy. For enterprises, effective governance provides the confidence to scale AI responsibly, balancing rapid adoption with ethical and regulatory imperatives.

2. CMPak AI Strategy

As a technology company, CMPak's AI strategy is anchored in rigorous AI research and governance-led design. Our vision is to build an AI-driven enterprise that is not only innovative and effective but also responsible, compliant, and resilient amid rapid technological change.

2.1. CMPak AI Research

CMPak is investing in R&D initiatives to track the evolution of generative AI globally and locally, running GenAI pilot scenarios across telecom operations, customer engagement, product catalog management, regulatory knowledge, DICT services, and B2B offerings. These efforts focus on designing AI-enabled workflows and strategically applying LLMs with external and domain-specific knowledge to improve operational efficiency, decision-making, and service innovation.

2.2. Alignment with Pakistan's National AI Policy

Pakistan's National AI Policy outlines a pillar-based framework to guide responsible AI adoption:

1. **Awareness & Readiness:** increasing public awareness, building applied research capacity, and reskilling the workforce through bootcamps and Massive Open Online Courses (MOOCs) Platforms.
2. **AI Market Enablement:** addressing ecosystem needs such as data standardization, accessibility, and computational resources.
3. **Progressive & Trusted Environment:** ensuring safe, ethical, and transparent use of AI with strong data privacy protections and regulatory oversight.
4. **Transformation & Evolution:** developing sector-specific roadmaps, enabling industrial transformation, and deploying regulatory sandboxes for controlled adoption.

CMPak's AI+ framework is designed to align with these national drivers. We embrace the goals of skills development, ethical governance, infrastructure readiness, and sectoral transformation while ensuring our AI initiatives not only comply with policy standards but also position CMPak as a key enabler of Pakistan's AI vision.

2.3. CMPak AI+ Framework

Built from our research insights and industry needs, the CMPak AI+ Framework provides the foundation for responsible, scalable, and enterprise-ready AI deployment. It is structured across three distinct layers:

- **Unified Infrastructure Foundation:** The unified architecture layer integrates computing resources (CPU/GPU), storage systems, and Data Lake, providing a scalable and efficient foundation for AI workloads.
- **Comprehensive AI Capabilities:** The Capability Layer provides a robust suite of modern AI technologies, spanning from foundational ML and NLP to advanced implementations like Reinforcement Learning, and Agentic AI, enabling sophisticated AI-driven solutions across the organization.
- **Business-Focused AI Application Layer:** The Application Layer demonstrates CMPak's strategic deployment of AI across critical business domains, including marketing optimization, network intelligence, management automation, DICT, and enhanced service delivery.

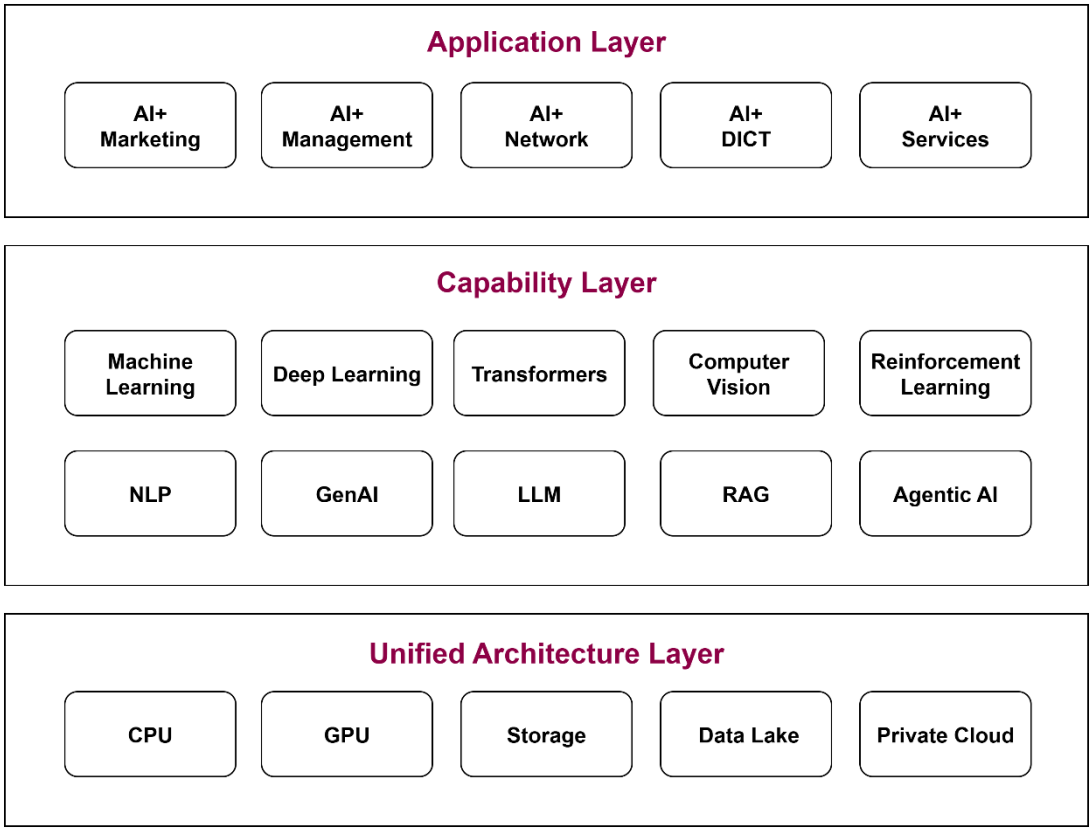


Figure: CMPak AI+ Framework

3. Challenges & Limitations: From Theory to Practice

The rapid growth of Large Language Models (LLMs) has created enormous excitement across industries, but their practical adoption in enterprises faces significant limitations. While organizations are quick to explore LLM-powered solutions, they often encounter challenges related to cost, control, and adaptability. Running and maintaining these models requires substantial compute resources, and enterprises struggle to align generic pre-trained capabilities with their domain-specific needs.

Another pressing issue is the problem of hallucinations and static knowledge. LLMs excel at generating human-like text, but they can produce inaccurate or fabricated information with confidence. This undermines trust, especially in compliance-driven environments such as telecom and finance, where reliability and factual accuracy are non-negotiable. Their knowledge also becomes outdated once the training cut-off is reached, making them unsuitable for domains where policies, regulations, or customer data change frequently.

Fine-tuning has traditionally been seen as a way to mitigate these shortcomings, but it introduces its own constraints. The process of gathering training data, curating it, and retraining or fine-tuning the model is expensive and time-consuming. More importantly, it cannot keep pace with environments where information evolves rapidly. In such dynamic contexts, fine-tuning becomes a bottleneck rather than a solution, limiting the model's adaptability.

Retrieval-Augmented Generation (RAG) has emerged as an organizational bridge to overcome these issues. By retrieving information from external sources at query time, RAG allows enterprises to ground model outputs in up-to-date, domain-specific knowledge. This not only reduces the dependence on repeated fine-tuning but also improves accuracy and flexibility in responding to complex queries. For industries dealing with heterogeneous documents, continuous updates, and regulatory oversight, RAG offers a more scalable and cost-effective approach.

Yet, RAG itself is not without challenges. Naive implementations often suffer from poor retrieval quality, fragmented or redundant context, and limited context windows within the model. These technical constraints can still lead to incomplete or misleading answers. Moreover, without a structured governance layer, enterprises face the risk of inconsistent responses, compliance gaps, and reduced transparency in how outputs are generated. RAG is also inherently passive; it can retrieve and generate information but cannot perform actions or trigger workflows. These shortcomings highlight the need for more advanced approaches that extend beyond static retrieval, setting the stage for Agentic AI as the next step in enterprise adoption.

4. Enterprise Knowledge Challenges

Telecom enterprises operate in a highly dynamic, compliance-driven environment where knowledge can be abundant, unstructured and frequently updated. Across an organization, vast information ecosystems exist such as regulatory frameworks, technical specifications, product portfolios, contractual obligations, operational procedures, and HR and procurement policies. These assets span multiple domains and reside in diverse unstructured formats.

While this information is technically accessible, practical usability is limited. People often struggle to locate accurate and contextually relevant content quickly because:

- **Volume & Fragmentation:** Policies and technical guidelines run hundreds of pages, distributed across multiple repositories.
- **Dynamic Change:** Telecom policies, compliance requirements, and product configurations change frequently, requiring constant updates to maintain accuracy.
- **Complexity & Terminology:** Domain-specific jargon, legal language, and cross-referenced rules make it difficult for non-experts to interpret and apply the right information.
- **Latency in Decision-Making:** Employees spend significant time searching, interpreting, and validating details, slowing operations and increasing the risk of missteps.

This is where **Natural Language Processing (NLP)** combined with **Retrieval-Augmented Generation (RAG)** provides a transformative advantage. RAG enables users to ask questions in natural language and receive precise, context-aware answers grounded in policy and process documents, rather than relying on manual document review. Instead of navigating complex file structures or searching across silos, users interact with a unified intelligence system that connects disparate knowledge, understands telecom context, stays current with changes, and delivers traceable, compliant answers.

By layering RAG on top of existing enterprise knowledge, CMPak is transforming static document repositories into dynamic, AI-powered knowledge systems. This shift empowers customer care teams to resolve queries instantly and accurately, enables HR to manage evolving internal processes with confidence, allows procurement to validate regulatory and contractual terms seamlessly, and supports compliance teams in maintaining oversight while driving faster, risk-aware decisions.

5. Enterprise-Grade RAG System Architecture

An enterprise-grade Retrieval-Augmented Generation (RAG) system requires rigorous design across all stages, from raw data ingestion to retrieval optimization, response synthesis, and continuous evaluation. CMPak's RAG solution is architected as a modular, multi-stage framework that ensures scalability, efficiency, and compliance while embedding advanced retrieval techniques to support complex enterprise knowledge use cases. It is designed to operate effectively across various functions within an organization.

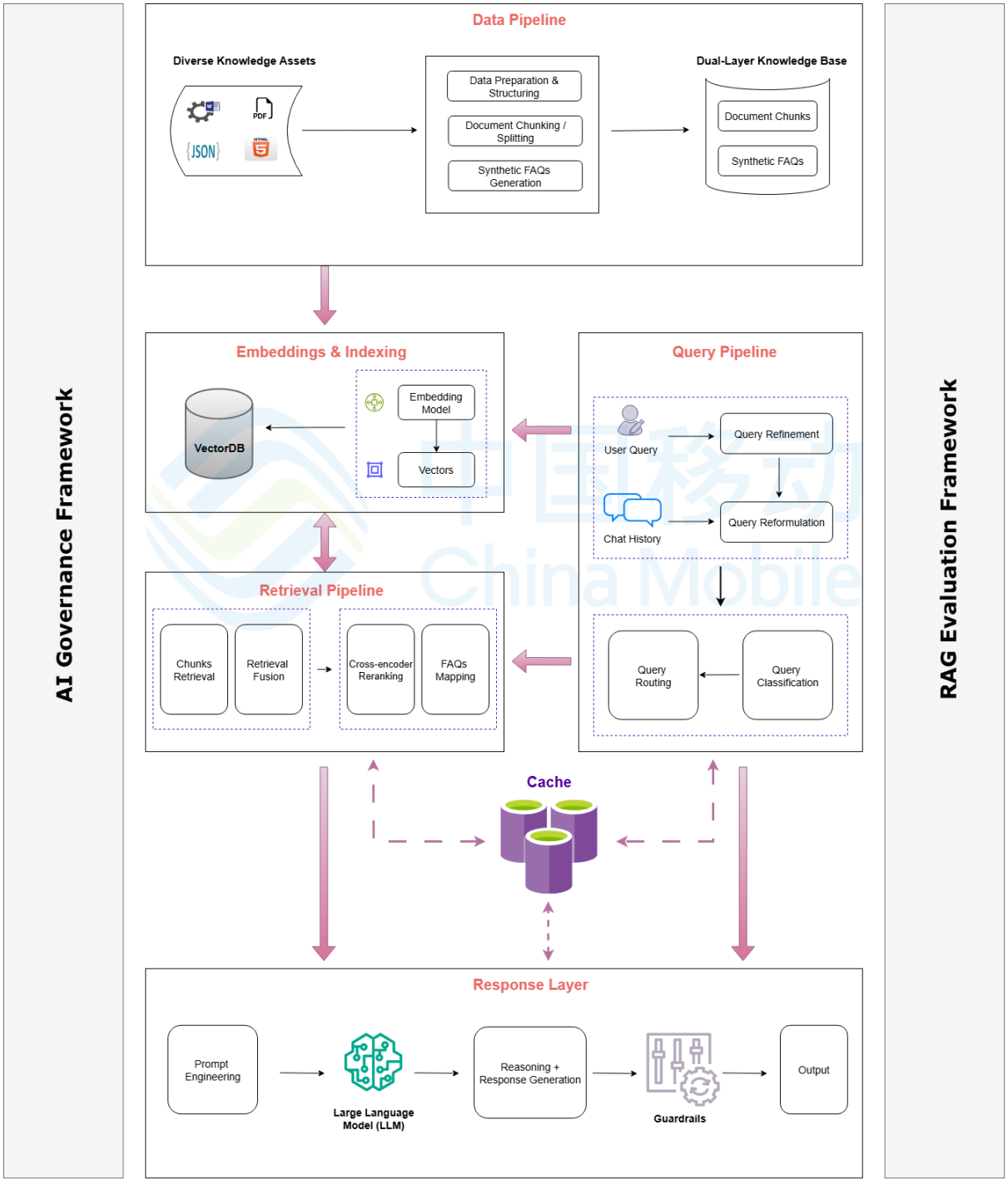


Figure: RAG System Architecture

5.1. Data Pipeline

Data Preparation and Structuring

Knowledge assets may span multiple domains and diverse formats, including PDFs, Word files, HTML pages, and internal portals, covering areas such as operations, HR, procurement, marketing, and IT. To ensure consistency, usability and machine-readability, a comprehensive standardization process is applied including:

- **Noise reduction:** Headers, footers, and decorative elements are removed to minimize irrelevant tokens and improve embedding quality.
- **Unified formatting:** Text content is normalized and converted into Markdown, providing a lightweight, structured format with clear delimiters and semantic cues.
- **Semantic segmentation:** Documents are split based on logical hierarchies such as headings and subheadings to maintain context integrity and avoid arbitrary breaks.
- **Sensitive data handling:** Personally identifiable and confidential information is automatically anonymized or removed where necessary to maintain compliance and privacy.
- **Metadata management:** Key attributes such as document type, department, creation date, and version history are extracted to enable better filtering, access control, and auditability.
- **Data normalization:** Duplicate and near-duplicate text blocks are detected and removed to reduce noise and improve retrieval efficiency.

This approach improves token efficiency, preserves context, and directly strengthens retrieval accuracy and the factual grounding of LLM responses.

Document Chunking / Splitting

A “structure-based chunking strategy” is applied across different document formats instead of using fixed-size windows, ensuring splits align with natural boundaries and preserve context. Examples include:

- **Markdown:** split by heading levels (H1, H2, H3)
- **HTML:** split by tag hierarchies such as sections or divs
- **JSON / YAML:** split by object or key-value structures

To support enterprise needs, additional steps include:

- **Chunk size optimization:** Maximum chunk length is tuned to balance token efficiency with contextual completeness.
- **Hierarchical linking and metadata:** Each chunk retains its parent section references and document metadata, ensuring traceability and enabling multi-level search.
- **Incremental updates:** Only modified sections are re-chunked when documents change, reducing reprocessing overhead.

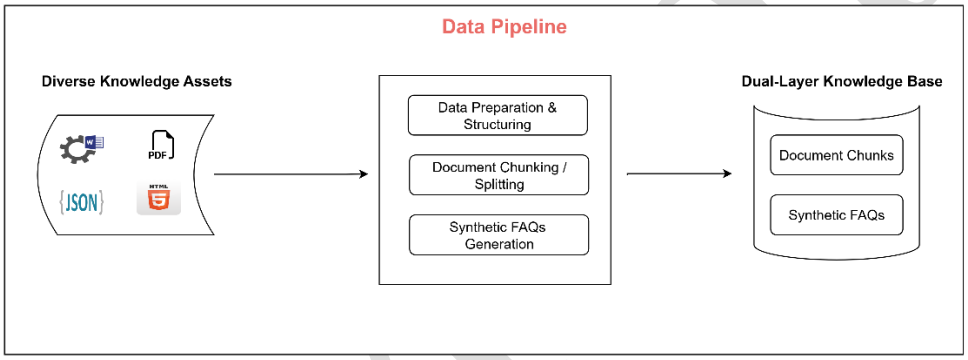
This approach produces self-contained, semantically coherent, and traceable chunks, while reducing the risk of fragmented context and improving retrieval precision and governance.

Synthetic FAQs Generation

To make the knowledge base more resilient to diverse and natural user queries, an "instruction-tuned LLM" is used to generate FAQs for each chunk.

- **Q/A generation:** Chunks are processed through prompt templates to create likely user questions paired with concise, accurate answers.
- **Traceable mapping:** Each generated Q&A pair is linked back to its source chunk through mapping dictionaries to maintain context and auditability.
- **Dynamic updates:** FAQs are automatically refreshed when source documents are modified, ensuring that generated questions and answers remain updated and consistent with the latest policies and procedures.
- **Multi-lingual support (where relevant):** FAQs can be generated in multiple languages (e.g., English and Urdu) to serve diverse user groups and enable seamless bilingual search.

This dual-layer knowledge base (documents + FAQs) expands query coverage, supports natural language search, and improves usability across domains such as customer care, HR, procurement, and compliance.



Document structuring, chunking strategy, and FAQ generation are adapted based on the target domain and use case requirements, ensuring that retrieval granularity, context depth, and question coverage align with the specific operational, compliance, or customer-facing objectives.

5.2. Embeddings and Indexing

Embeddings

A core component of RAG is the embedding layer, where text is transformed into numerical vectors that capture semantic meaning. By representing words, phrases, or entire passages as vectors, embeddings act as the bridge between user queries and the knowledge base, enabling the system to surface content that is semantically aligned with the intent of the request.

Multiple embedding models are evaluated for semantic similarity, retrieval accuracy, latency, and multilingual performance across domain-specific use cases:

Category	Models	Primary Advantage
English-focused	BAAI/bge-small-en-v1.5 BAAI/bge-large-en-v1.5	High semantic precision for English corpora

Multilingual	intfloat/multilingual-e5-large intfloat/multilingual-e5-small sentence-transformers/LaBSE	Strong multilingual & cross-lingual retrieval
Lightweight	sentence-transformers/all-MiniLM-L6-v2 sentence-transformers/all-MiniLM-L12-v2 sentence-transformers/all-mpnet-base-v2 distilbert/distilroberta-base	Fast inference and low compute cost
QA-optimized	sentence-transformers/multi-qa-MiniLM-L6-cos-v1	Tuned for question–answer retrieval tasks

When general-purpose embedding models do not fully capture telecom and regulatory language, **domain-specific fine-tuning** is considered.

Vector Database

Once embeddings are generated, they are indexed in a vector database (VectorDB)**,** which enables efficient similarity search across large collections of documents. Unlike relational databases built for structured queries, VectorDBs are optimized for high-dimensional vector comparisons, which are essential for fast and relevant retrieval in RAG systems. Several VectorDB options are evaluated based on use case, latency, scalability, memory efficiency, and integration flexibility:

Database	Key Features	Best Fit
Chroma	Open-source, Python-native, seamless integration with LangChain/LlamaIndex	Rapid prototyping and LLM app development
Pinecone	Managed service, elastic scaling, low-latency search	Production-grade, cloud-first deployments
Weaviate	Hybrid (dense + sparse) retrieval, API-driven, metadata filtering	Enterprise search with structured filtering
FAISS	GPU-accelerated, optimized for large-scale ANN search	High-performance and research environments
Qdrant	Open-source, payload filtering	Vector search with structured metadata
Milvus	Distributed, scalability, multi-tenant secure	Large-scale enterprise AI search
PGVector	PostgreSQL extension, SQL-based ANN queries	Adding vector search to existing relational systems

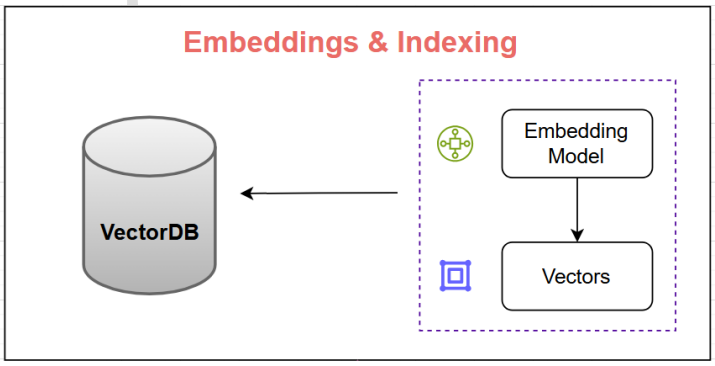


Figure: Embeddings & Indexing

5.3. Query Pipeline

Query Refinement

User queries often contain ambiguity or informal phrasing. To improve precision, inputs are pre-processed to remove noise and normalized into semantic-friendly representations suitable for embedding models.

Query Reformulation

In a conversational RAG setting, queries are frequently ambiguous or dependent on prior dialogue. To address this, lightweight LLMs are used to reformulate user inputs using chat history, converting them into clear, self-contained queries. This process incorporates conversational context, resolves references, and reduces ambiguity, ensuring retrieval aligns with the broader dialogue.

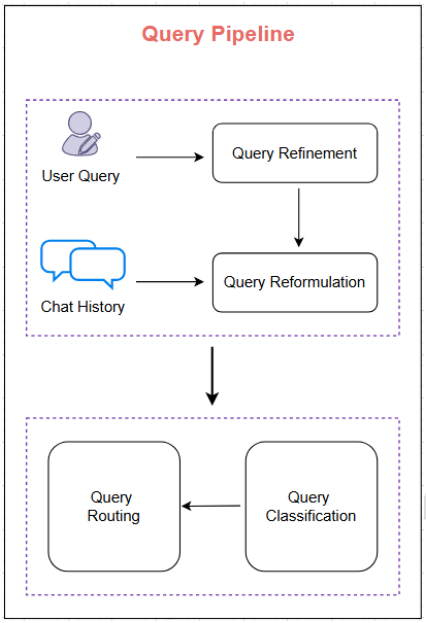
Query Intent Classification and Routing

Not all user queries require the same retrieval strategy. To optimize performance and accuracy, an intent classification layer is applied before retrieval. This step ensures that queries are routed based on their underlying purpose rather than treated uniformly. By categorizing inputs, the system avoids unnecessary retrieval, reduces latency, and improves contextual relevance.

Through experimentation, a set of intent-specific strategies has been identified and is now applied across multiple domains.

Query Intent Type	Characteristics	Routing Strategy	Example Query	Typical Domains
Fact Lookup	Direct, single-hop queries needing precise information	Dense + sparse retrieval with reranking	"What is the franchise settlement policy?"	Sales & Distribution, Compliance
Procedural	Step-by-step, workflow-aligned instructions	Retrieval + workflow templates	"How do I apply for approval in system?"	HR, Procurement, Operations
Reasoning-Heavy	Multi-hop, contextual, inference-based	Multi-hop retrieval + iterative reranking	"Compare settlement policies for franchises and retailers."	Regulatory Affairs, Legal, Policy Analysis
Non-Retrieval / Conversational	Greetings, acknowledgments, or chit-chat	Direct response generation (no retrieval)	"Hello" / "Thanks for the help"	Customer Care, General Use

Frameworks such as **LangChain** and **LangGraph** provide orchestration capabilities that enable this routing. They allow dynamic selection of retrieval methods, reranking when required, and direct response generation for low-information intents. This approach keeps the system efficient, reduces retrieval overhead, and ensures responses are contextually aligned with user intent.



5.4. Retrieval Pipeline

Chunks Retrieval

Once queries are reformulated and intent-classified, they are converted into embeddings and compared against stored chunk vectors in the vector database. To capture both conceptual meaning and domain-specific terminology, a "hybrid retrieval strategy" is used in most enterprise scenarios:

- **Dense retrieval (vector-based):** Retrieves semantically related chunks from a vector database (VectorDB) to improve contextual relevance.
- **Sparse retrieval (BM25/keyword-based):** Ensures lexical precision and captures exact terms, such as regulatory codes and telecom-specific jargon.

Retrieval depth is tuned with various top-k settings (1–20) to balance high recall (broad coverage) with low latency (faster response times).

Retrieval Fusion

Dense and sparse retrieval outputs are consolidated through a "fusion layer" that assigns weighted rankings, combining semantic relevance from dense models with lexical accuracy from sparse models.

- Weights are tuned experimentally for each domain to reflect its criticality. For example, legal and compliance documents require high lexical precision, while operational policies benefit from semantic coverage.
- Low-confidence or duplicate results are filtered before reranking, improving precision and reducing noise.

Cross-Encoders Reranking

Because LLMs have finite context windows, only a limited number of chunks can be passed for response generation. A cross-encoder re-ranking layer ensures that the most relevant chunks are selected. Unlike bi-encoders that compute embeddings separately, cross-encoders jointly encode the query and chunk, producing a direct relevance score.

- This step significantly improves precision, surfacing the most contextually aligned chunks.
- Typically, the top 1 to 5 re-ranked chunks (domain-dependent) are forwarded to the LLM for response synthesis, ensuring factual accuracy without exceeding token limits.

Different models for reranking:

Model	Key Traits	Best Fit
cross-encoder/ms-marco-MiniLM-L6-v2	Very lightweight, fast inference	Low-latency, smaller workloads
cross-encoder/ms-marco-MiniLM-L12-v2	Lightweight, strong chunk ranking	Production-ready, high accuracy
tomaarsen/reranker-distilroberta-base-nli	Distilled, low-latency	Enterprise, latency-sensitive
antoinelouis/crossencoder-electra-base-mmarcoFR	Balanced size vs. performance	Mid-scale deployments

FAQs Mapping

Because the knowledge base includes synthetic FAQs linked back to original document chunks, after re-ranking the system performs "FAQ-to-source consolidation". If multiple highly ranked FAQ entries originate from the same underlying chunk, they are merged and mapped back to that single source. This reduces redundancy, and ensures the LLM receives compact, non-repetitive context.

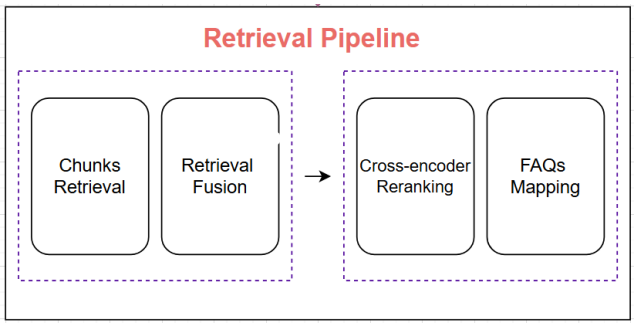
Whether FAQ chunks are retained as separate entries or consolidated back to their original document chunks is determined by the requirements of the specific use case and domain. In most cases, consolidating to the original chunk provides richer context to the LLM for generating more accurate and reliable responses.

Caching

To optimize latency and reduce redundant computation, frequent queries and their re-ranked top chunks are cached. This creates a "hot cache" of high-demand queries and avoids executing the full retrieval and re-ranking pipeline repeatedly.

- **Vector-level caching:** Embedding search results and their re-ranked lists are stored in a low-latency store. Common queries can be served directly from this layer.
- **Response-level caching:** Final LLM outputs for deterministic or policy-stable questions are stored, allowing instant reuse.

Caches use time-to-live (TTL) controls, invalidation hooks, and content hashing to ensure updates are automatically applied when source documents change, maintaining compliance and accuracy.



5.5. Response Layer

Prompt Engineering

In conversational RAG, prompt engineering determines how effectively selected chunks are transformed into reliable outputs. Prompts include fallback strategies that instruct the model to acknowledge uncertainty when retrieval is incomplete, reducing unsupported or fabricated answers. Frameworks such as LangChain provide a variety of pre-built prompt templates for tasks like question answering, summarization, classification, and conversational flows. These templates are further customized to meet enterprise-grade requirements, reinforcing compliance, domain-specific accuracy, and consistency across CMPak's use cases.

Model Selection

Selecting the right LLM is critical because no single model performs optimally across all enterprise scenarios. Different use cases demand different trade-offs: complex reasoning tasks require models with deeper capacity and larger context windows, while high-volume or latency-sensitive applications benefit from smaller, faster models. Factors such as domain-specific vocabulary support, integration with retrieval pipelines, compliance with internal governance rules, and infrastructure constraints also influence which model is most effective for a use case. Benchmarking across these dimensions ensures that each deployment uses an LLM aligned with its accuracy, cost, and performance requirements.

Example LLM options for enterprise deployment are:

Model	Key Traits	Use Cases	Deployment Alignment
DeepSeek-R1 (1.5B-7B)	Optimized reasoning with lower compute overhead	Procurement analysis, compliance-heavy use cases	Moderate GPU footprint, enterprise servers
Qwen (1.5B–72B)	Balanced accuracy and efficiency, strong instruction following	Customer service, multilingual product support, knowledge assistants	Flexible scaling: GPU clusters or hybrid cloud
LLaMA (3B–70B)	Strong reasoning, multilingual capability, open-source ecosystem	Policy compliance checks, regulatory queries, HR knowledge	GPU-heavy, data center deployment
Mistral (7B/8x7B MoE)	High performance, efficient inference, strong open-source adoption	Real-time customer care, high-throughput multi-turn dialogues	Distributed deployment with MoE optimization
JIUTIAN-139MoE-Chat	Strong on industrial & general benchmark	Industry-specific reasoning (telecom, energy, finance, medical), compliance-heavy tasks	Scales on GPU/NPU clusters, suited for secure enterprise deployments
Distilled Models	Low-latency, lightweight deployment	Latency-sensitive services, high-volume query processing	Edge devices, CPU/GPU-constrained environments

Response Generation

The prompted input combining the user query, relevant context, and conversational history is passed to the selected LLM to generate coherent, contextually accurate responses. Hyperparameters are tuned to balance accuracy, efficiency, and adaptability across use cases:

- **Temperature** controls the balance between deterministic and variable outputs.
- **Sampling strategies** regulate response diversity while preserving factual grounding.
- **Context window size** is adjusted to ensure sufficient coverage of retrieved content while optimizing computational cost.

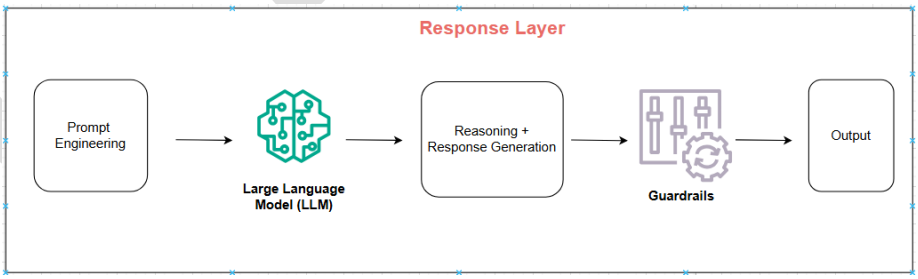
Careful configuration ensures generated output remains fluent, accurate, and domain aware.

Guardrails & Output Validation

Even with strong prompt engineering and careful model selection, generated outputs require an additional layer of validation before delivery. Post-generation guardrails act as a safeguard, ensuring responses are accurate, safe, and compliant with enterprise and regulatory requirements. Key mechanisms include:

Guardrail Type	Objective	Example Applications
Policy Compliance Filters	Block or flag outputs that violate regulatory, telecom, or HR guidelines.	Prevent disclosure of restricted policies or pricing data.
Toxicity & Personal Identification Detection	Detect and remove harmful language or personally identifiable information.	Screen customer care responses for sensitive data leaks.
Stylistic Enforcement	Enforce organizational tone, structure, and response length.	Maintain concise, professional answers in HR assistants.
Escalation to Human-in-the-Loop (HITL)	Route high-risk or ambiguous outputs to domain experts for final approval.	Procurement AI agent escalating vendor recommendations.

This guardrail layer operationalizes governance-by-design, reinforcing trust, accountability, and compliance across all AI-powered interactions.



5.6. Evaluation Framework

Ensuring robustness, compliance, and enterprise readiness requires a systematic approach to test every RAG use case before deployment. A comprehensive evaluation framework validates each stage of the pipeline including embedding, routing, retrieval, re-ranking, prompting, and response generation to ensure accuracy, efficiency, and reliability. The process follows a structured multi-tier design.

Supervised Testing

Each new use case is evaluated against curated, domain-specific test sets that reflect real-world user queries. The framework runs supervised tests at every pipeline stage to detect bottlenecks, failure points, and accuracy gaps early.

Quantitative Evaluation

Automated metrics validate retrieval and generation quality. Typical measures include:

- **Recall@k and Precision:** retrieval accuracy and coverage
- **Mean Reciprocal Rank (MRR):** ranking quality
- **Latency and Throughput:** system responsiveness under load
- **Context Coverage:** proportion of relevant material surfaced to the LLM

Metrics are dynamically prioritized depending on the domain (e.g., recall and precision for compliance vs. latency for customer care).

Qualitative Evaluation

Outputs are assessed through side-by-side analysis of strong vs. weak responses, scored on factual correctness, completeness, clarity, and tone adherence. This step ensures the system produces outputs aligned with organizational communication standards.

Human-in-the-Loop (HITL) Feedback

For high-stakes domains (e.g., legal, compliance, procurement), outputs undergo expert review in testing phase. This step adds accountability and regulatory assurance while creating continuous feedback loops to refine prompt templates, routing logic, and retrieval strategies.

Baseline Comparison

RAG outputs are systematically benchmarked against fine-tuned LLMs and alternative hybrid approaches for selected use cases. This evaluation validates that the RAG architecture maintains high adaptability and cost efficiency while preserving accuracy and domain reliability.

Domain-Specific Evaluation Metrics

Domain	Priority Metrics	Rationale
Compliance / Regulatory	Recall@k, Precision, Accuracy	Critical to retrieve every relevant clause and avoid policy omissions.
HR & Internal Policies	Consistency, Fairness, Clarity	Ensures equitable and unambiguous responses aligned with organizational policies.
Procurement	Relevance, Completeness, Transparency	Supports vendor comparison, contract drafting, and compliance with governance rules.
Customer Care	Latency, Fluency, Recall@k, Multilinguality	Fast and clear responses are essential for high-volume, customer-facing interactions.
DICT / B2B Solutions	Adaptability, Accuracy, Multilinguality	Reliable and language-aware outputs for enterprise clients.

Configuration Benchmarking

To achieve optimal domain alignment, the framework systematically benchmarks and compares alternative configurations, including:

- Embedding models (e.g., multilingual vs. domain-adapted)
- Query reformulation and routing strategies
- Dense vs. sparse retrieval strategies and hybrid fusion methods
- Alternative cross-encoder re-rankers
- Multiple prompt templates (e.g., conversational vs. formal)
- Candidate LLMs varying in size, latency, reasoning depth, and domain fit

The best-performing setup is selected per use case, ensuring each deployment balances accuracy, speed, and cost for its intended domain (e.g., HR, procurement, compliance, customer care).

This evaluation framework enables continuous testing, new updates or models are validated against historical baselines to ensure no degradation in quality. Together, automated testing, domain-driven metrics, and expert validation establish a governance-first approach that ensures safe, accurate, and scalable AI deployment.

6. Agentic AI Approaches

The transition from Large Language Models (LLMs) to Retrieval-Augmented Generation (RAG) has already made enterprise AI systems far more reliable and contextually grounded. However, RAG pipelines remain fundamentally reactive: they respond to prompts with accurate context but do not plan actions, adapt workflows, or interact autonomously with enterprise systems. Agentic AI represents the next stage in this evolution, transforming AI from a static information retrieval engine into a decision-support and action-execution system.

Agentic AI enables models not just to retrieve and answer but also to plan, reason, and act. AI agents autonomously break down complex user goals into actionable sub-tasks, orchestrate enterprise APIs and data sources, apply tool-augmented reasoning, and execute multi-step workflows, all while ensuring governance and compliance. This shifts AI from being a passive assistant into a proactive operator that coordinates knowledge, tools, and business processes safely and efficiently.

6.1. Extending RAG into Agentic Capabilities

Enterprises are extending their RAG pipelines into agentic systems with the following key enhancements:

- **Dynamic Retrieval Strategy Selection:** Agents autonomously pick retrieval strategies based on the query, switching between dense, sparse, or hybrid approaches as required. For example, compliance-heavy queries leverage from keyword-based precision, while customer care queries benefit from semantic similarity search, ensuring both accuracy and efficiency.
- **Governance Aware Execution:** Governance-by-design is embedded into every step of the agent workflow. Access permissions, data privacy checks, and audit logging are enforced natively. Agents cannot trigger actions outside approved boundaries or access non-whitelisted tools.
- **Adaptability with Compliance Controls:** Agents integrate directly with enterprise APIs to trigger workflows, such as updating HR cases, generating RFP drafts, or analysing regulatory changes while remaining inside compliance guardrails that define safe operational limits.

6.2. Agentic AI Frameworks and Architecture

Building enterprise-ready agentic AI requires a layered, governance-first architecture rather than simply extending LLM capabilities. A robust architecture typically includes:

1. **Core LLMs**: Models capable of long-context reasoning, multi-step planning, and controlled tool invocation to execute complex enterprise tasks.
2. **Knowledge Layer**: Conversational RAG pipelines supplying live, context-rich, and domain-aligned information to ground decisions.

3. **AI Agents & Tool Interfaces**: Secure connections with tool interfaces including APIs, databases, and enterprise platforms (CRM, HRMS, ERP, compliance systems) are used in order to enable AI agents to perform actions.
4. **Governance Modules**: Embedded access controls, audit logs, approval workflows, and compliance checkpoints to ensure safe and accountable execution.
5. **Memory & State Management**: Persistent and ephemeral memory to maintain context across long workflows while preventing sensitive data leakage, enabling continuity without sacrificing security.
6. **Reasoning Trace**: Built-in reasoning logs and decision traces to provide transparency, support audits, and allow human oversight for high-stakes actions.
7. **Observation & Feedback Loops**: Integrate reinforcement learning with human feedback (RLHF) or automated evaluation signals to continuously refine agent reasoning and decision-making.
8. **Orchestration Layer**: Agent frameworks (e.g., LangGraph, LangChain) coordinating multi-agent collaboration, task decomposition, and dynamic retrieval strategy selection.

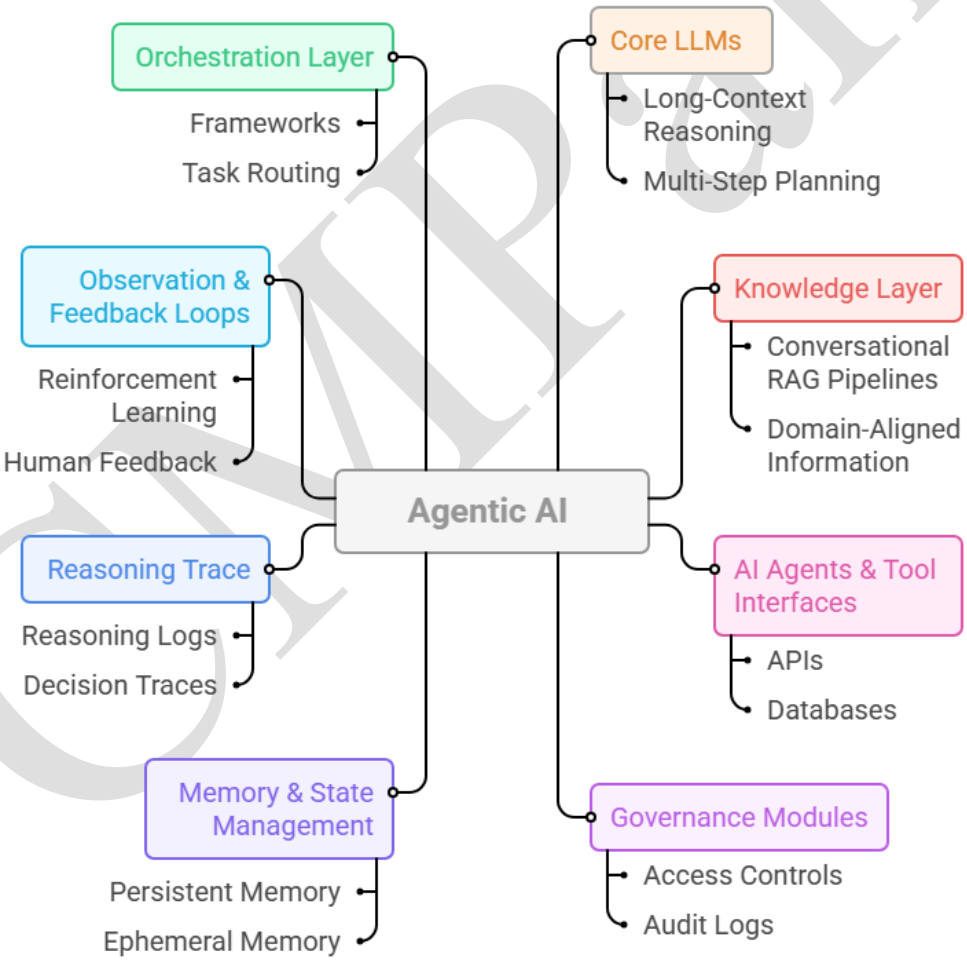


Figure: Agentic AI Framework

6.3. Use Cases and Applications of Agentic AI

Agentic AI unlocks a wide range of opportunities across CMPak and the broader telecom sector:

Use Case	Description
Policy & Compliance Agents	Interpret evolving regulations, monitor policy changes, and draft compliance report.
Procurement Agents	AI agents create RFP/RFI documents, perform vendor analysis, provide recommendations while referencing governance rules, and can even engage with potential vendors to gather information or initiate collaboration.
Customer Service Agents	Manage multi-user support, escalate complex issues, and auto-log cases into CRM systems.
HR Assistants	Answer employee queries, draft HR responses aligned with policy, and escalate sensitive cases when required.
DICT & B2B AI Services	Offer enterprise clients AI-driven insights and custom solutions that can reason across their business data.
Network Operations Agents	Self-intelligent autonomous network L4 AI agents provide real-time troubleshooting for field engineers by analysing historical incidents, network topology, and technical manuals and perform autonomous AI-driven network management.

6.4. Future Outlook: Embedded AI Orchestration

The future of enterprise AI lies in embedding Agentic AI into core business workflows. Instead of existing as stand-alone assistants, agents will become integrated into everyday operations, drafting regulatory filings, managing procurement cycles, supporting B2B client services, and running self-intelligence autonomous network.

For CMPak, this means moving from isolated pilots to enterprise-wide orchestration**,** where multiple specialized agents collaborate under shared governance frameworks. Such integration will reduce manual effort, increase accuracy in compliance-heavy functions, and create new opportunities for B2B product innovation.

7. Governance Framework for AI Solutions

CMPak adopts a Governance-by-Design approach, embedding risk controls, accountability, and compliance directly into the architecture and development lifecycle of all AI systems. By making governance an integral design principle rather than a post-deployment checkpoint, AI deployments remain transparent, secure, fair, and regulatory compliant at scale.

7.1. Governance-by-Design Overview

Governance-by-Design is a proactive methodology that integrates governance principles throughout the lifecycle of an AI system. It shifts governance from being a compliance checkpoint to becoming a fundamental design characteristic. The guiding principles of this approach include:

- **Transparency**: Clear communication of the AI system's capabilities, limitations, and operational processes.
- **Fairness**: Identification and mitigation of potential bias to ensure equitable outcomes across use cases.
- **Explainability and Interpretability (XAI)**: Providing understandable documentation and justifications for the system's decisions and outputs.
- **Accountability**: Defining clear ownership of responsibilities across development, deployment, and oversight.
- **Privacy and Data Governance**: Safeguarding personal information and ensuring that data handling practices are secure, ethical, and compliant with applicable regulations.

These principles are operationalized through three governance pillars: AI Governance, Data Governance, and Process Governance.

7.1.1. AI Governance

AI Governance addresses model behavior, outputs, and ethical implications, ensuring that all systems align with CMPak's governance principles of transparency, fairness, explainability, accountability, and privacy.

Model Risk Management Across the AI Lifecycle

Governance is applied throughout a structured AI lifecycle to ensure responsible outcomes:

- **Planning**: Define the use case, success metrics, and potential risks upfront, ensuring alignment with business objectives and ethical guidelines.
- **Data Collection and Pre-processing**: Govern data ingestion and curation with emphasis on quality, provenance, and bias detection, supporting fairness and privacy.
- **Model Building and Interpretation**: Ensure explainability for interpreting decisions, supported by artifacts at different stages of the AI lifecycle documenting rationale & performance across segments.
- **Post Validation**: Continuously test models against benchmarks for accuracy, fairness, robustness, and security.

Governance Artifacts

Each deployed solution is accompanied by a comprehensive documentation that enforces governance principles and provides traceability. Key fields include:

- **Model Description and Intended Use**: Clarifies purpose and application scope.
- **Training Data and Sources**: Records dataset origin and sourcing details (internal, open source, etc.), sensitivity level of data undergoing training as well as the preprocessing steps taken on data to ensure fairness & privacy.
- **Evaluation Metrics**: Reports performance of model by documenting test dataset details, rationale for choice of evaluation model, and key performance results from testing alongside rationale for continuing with results.
- **Ethical Considerations and Risks**: Identifies biases, potential failure modes, and broader impacts.
- **Hardware and Infrastructure**: Specifies deployment requirements.
- **Decision-Making and Oversight**: Defines ownership and human review mechanisms.

Responsible AI Use

Human-in-the-Loop (HITL) controls are embedded in high-stakes decision areas to ensure that accountability remains with humans. This mitigates automation bias and positions AI as an augmentation tool rather than a replacement for responsible decision-making.

7.1.2. Data Governance

The Data Governance pillar ensures that all information used and generated by AI systems remains accurate, secure, and compliant, operationalizing the principle of Privacy and Data Governance.

Data Lifecycle Management: Dedicated teams oversee the full data lifecycle, including source validation, handling of personally identifiable information (PII), periodic recertification of knowledge sources, and continuous quality checks to maintain the reliability of RAG systems.

Security and Compliance: Strict Role-Based Access Control (RBAC), encryption, and data sovereignty measures are applied to safeguard sensitive information and ensure adherence to regulatory requirements.

7.1.3. Process Governance

Process Governance ensures that AI systems are developed, modified, and monitored under structured and auditable practices.

AI-Specific SDLC: Governance checkpoints are integrated at each stage of the AI lifecycle for reinforcing transparency and accountability.

Change Control: A formal Change Management policy requires approval from designated Change Approval Authority for significant modifications to models, datasets, or prompts, ensuring oversight and risk mitigation.

Assurance and Monitoring: Continuous performance monitoring, supported by audit trails of AI interactions, provides accountability and enables ongoing improvements.

7.2. Governance for Agentic AI Workflows

Agentic AI introduces autonomy into enterprise systems, requiring enhanced safeguards to ensure that actions remain accountable, explainable, and compliant.

- **Action Safeguards**: High-impact actions such as sending external communications, modifying customer records, or initiating payments trigger a pause in the workflow and require explicit Human-in-the-Loop (HITL) approval. This ensures accountability for irreversible outcomes.
- **Tool and API Allow-Lists**: Agents are restricted to a catalog of approved internal and external APIs. Each tool is assessed for risk classification (e.g., read-only, data-writing, or financial transaction) to prevent unauthorized or unsafe use.
- **Reasoning Transparency**: For every deployed agent, reasoning steps are documented and linked to the Model Card. A designated owner is responsible for ensuring that decision-making logic remains interpretable and explainable.
- **Context Window and Step Limits**: Agents are bounded by predefined limits on reasoning steps and context size. These controls prevent runaway loops, resource exhaustion, and unbounded task execution.
- **Cross-Boundary Data Flow Monitoring**: Data flows between tools and knowledge sources are actively monitored to ensure that sensitive information is not inappropriately transferred across systems, maintaining privacy and regulatory compliance.

7.3. Governance Controls Matrix

Governance Layer	Control Objective	Example Control Mechanism
AI Governance	Ensure Fairness & Explainability	• Artifacts' maintenance • Bias testing across data slices • Documentation of reasoning, evaluation factors, and trade-offs
	Maintain Accountability	• Mandatory HITL for high-risk actions • Defined model owner and developer • Restricted use of whitelisted tools and APIs
Data Governance	Uphold Privacy & Security	• Compliance with customer data protection regulations • Enforcement of data sovereignty requirements
Process Governance	Ensure Reproducibility & Audit	• Version control for models, data, and prompts • Formal approval processes for all significant changes

This structured framework provides enterprises with a practical path to harness Generative AI effectively, embedding accountability, security, and compliance as foundational design principles rather than post-deployment checks. CMPak applies this matrix as an operational guide, ensuring that every AI deployment remains transparent, well-governed, and aligned with regulatory requirements.

8. Key Takeaways

The following are key considerations for designing and developing Retrieval-Augmented Generation (RAG) and Agentic AI systems that are scalable, compliant, and high-performing:

Importance of Data Quality

Enterprise AI strength depends on robust data governance and semantic structuring. Beyond basic cleansing, applying domain-aware normalization, metadata enrichment, and automated personal identification handling ensures that knowledge bases remain compliant and retrieval-ready.

Domain-Aware Embedding Strategy

Different data modalities such as policies, contractual clauses, FAQs, technical manuals require embedding models tuned for domain semantics and multilingual needs. Standardized selection and continuous evaluation maintain retrieval consistency across functions like HR, procurement, and compliance.

Hybrid & Adaptive Retrieval

Static retrieval pipelines are insufficient. Combining dense + sparse + hybrid search with dynamic top-k tuning and context-sensitive query rewriting significantly boosts accuracy and recall, especially in compliance-heavy and regulated environments.

Continuous & Automated Evaluation

Automated RAG testing frameworks should not be one-off. Continuous benchmarking of embedding drift, retrieval latency, and factuality ensures the system adapts to knowledge updates and model changes without degrading performance.

Agentic AI for Dynamic Workflows

Agentic AI reduces manual effort by autonomously planning and executing multi-step tasks while adhering to compliance rules. It adapts its actions to policy constraints and automatically escalates cases to human review whenever predefined risk thresholds are reached.

Human Feedback as a Governance Mechanism

HITL (Human-in-the-Loop) review is essential for high-stakes outputs such as procurement decisions, compliance checks, or regulatory responses. This approach also powers Reinforcement Learning from Human Feedback (RLHF) or preference optimization for domain-specific alignment.

Governance Frameworks

Governance-by-Design is essential; it serves as the control center for AI systems. Data quality frameworks, audit logs, and access controls keep AI trustworthy, and compliant.

9. Future Work

CMPak's next phase of Generative AI research and development focuses on advancing beyond RAG and Agentic AI foundations toward more autonomous, explainable, and enterprise-scale systems. The roadmap prioritizes directions in advanced reasoning, domain specialization, efficient scaling, and stronger governance integration.

9.1. Research Directions

Neural-Symbolic Hybrid Reasoning

Future agentic workflows will integrate symbolic governance rules with neural reasoning, creating hybrid agents that can both plan autonomously and enforce hard compliance constraints. This approach promises explainability, verifiability, and stricter adherence to regulatory requirements in telecom and financial domains.

Multi-Agent Consensus and Adjudication

For high-stakes use cases, multiple agents can be deployed to generate, critique, and refine outputs through debate or consensus before surfacing a final response. This approach reduces the risk of single-agent bias or error and is particularly suited for compliance-heavy workflows in HR, procurement, and legal domains.

Domain-Specialized Reranking with Mixture-of-Experts (MoE)

Reranking quality will be enhanced through domain-specialized re-rankers, orchestrated via a Mixture-of-Experts (MoE) architecture. Queries will be routed dynamically to the most suitable model, improving accuracy while optimizing compute usage. Sparse activation MoEs will be evaluated to achieve scale without prohibitive cost.

Voice-Enabled RAG with Secure Speech Interfaces

Integrating Automatic Speech Recognition (ASR) with RAG will enable real-time voice-based interactions for customers and employees. Research will focus on fine-tuning ASR for telecom-specific speech, multi-speaker diarization, and integrating voice biometrics for identity assurance, combining natural conversation with secure governance.

9.2. Enterprise-Scale Optimizations

Domain-Specific Embedding Fine-Tuning

Embedding models fine-tuned on CMPak's proprietary corpora, including policies, customer care scripts, and regulatory documents, will improve semantic granularity and retrieval consistency. Approaches include contrastive learning with hard negatives, adapter-based fine-tuning for efficiency, and distillation into smaller models for real-time use.

Distilled Cross-Encoders

High-performing but resource-intensive cross-encoders will be compressed through knowledge distillation. Techniques under review include layer pruning, attention head reduction, and teacher–student transfer.

Synthetic QA Generation at Scale

Synthetic Q&A pipelines will be scaled to enrich retrieval training and evaluation datasets. Future iterations will include adversarial QA generation to stress-test systems, multi-lingual QA generation for Pakistan's diverse user base, and automated validation loops to maintain compliance integrity.

Multi-Agent AI Workflow Orchestration

Future deployments will embed agents more deeply into enterprise systems with hierarchical coordination and compliance-aware planning. Supervisory agents will manage specialized sub-agents across CRM, ERP, and HRMS integrations, enabling secure orchestration of end-to-end business workflows.

10. Conclusion

The journey from Large Language Models (LLMs) to Retrieval-Augmented Generation (RAG) and now Agentic AI reflects how enterprise AI is evolving from static experimentation to dynamic, production-grade systems. RAG has improved accuracy, accessibility, and compliance by grounding responses in authoritative knowledge, while Agentic AI adds adaptability through reasoning, multi-step workflows, and secure tool integration.

The effectiveness of these systems depends on governance-by-design by embedding transparency, accountability, and regulatory alignment at every layer rather than treating compliance as an afterthought. This approach enables scalability, risk control, and trustworthy AI outcomes.

Looking ahead, advances in domain-specific embeddings, intelligent reranking, and multi-agent orchestration will further increase impact. By investing in these directions and maintaining strong governance, CMPak is positioned to drive measurable business value, strengthen compliance and decision-making, and continue shaping responsible AI adoption across the telecom sector.

As a technology-oriented company, CMPak is expanding its AI capabilities not only to transform operations but also to empower its B2B clients. By offering secure, enterprise-grade AI services, CMPak enables businesses to innovate faster, reduce operational risk, and unlock new growth opportunities. This focus on operational excellence and customer enablement reinforces CMPak's role as a trusted technology partner driving AI maturity across the industry.