TÍCH HỢP **OLLAMA** VÀO HỆ THỐNG **EOFFICE** (MICROSERVICES + RAG + IAM)

Contents

[1 — Tóm tắt chiến lược 2](#_Toc208577788)

[2 — Nhân sự & vai trò (gợi ý) 2](#_Toc208577789)

[3 — Giai đoạn & timeline ước lượng (tham khảo) 2](#_Toc208577790)

[4 — Giai đoạn 1: Discovery & Planning (2–3 tuần) 2](#_Toc208577791)

[5 — Giai đoạn 2: POC — RAG + Ollama (4–6 tuần) 3](#_Toc208577792)

[6 — Giai đoạn 3: Pilot (6–8 tuần) 4](#_Toc208577793)

[7 — Giai đoạn 4: Rollout toàn công ty (8–12 tuần) 5](#_Toc208577794)

[8 — Giai đoạn 5: Vận hành & Tối ưu (ongoing) 5](#_Toc208577795)

[9 — Hạ tầng & sizing tham khảo 6](#_Toc208577796)

[10 — Bảo mật & Quyền riêng tư (must-have) 6](#_Toc208577797)

[11 — RAG engineering (best practices) 6](#_Toc208577798)

[12 — Testing & QA 7](#_Toc208577799)

[13 — Rollback & Contingency 7](#_Toc208577800)

[14 — KPIs & Success Metrics 7](#_Toc208577801)

[15 — Chi phí & kinh phí (rất sơ bộ) 7](#_Toc208577802)

[16 — Ví dụ kỹ thuật nhanh (có thể dùng ngay) 8](#_Toc208577803)

[16.1 — Ví dụ Modelfile (ví dụ tối giản) 8](#_Toc208577804)

[16.2 — Mẫu Prompt Template (RAG) 8](#_Toc208577805)

[17 — Rủi ro chính & mitigation (ngắn gọn) 9](#_Toc208577806)

[18 — Acceptance checklist trước khi mở rộng 9](#_Toc208577807)

[19 — Gantt Timeline (ước lượng ~6 tháng) 9](#_Toc208577808)

[Giai đoạn 1 — Discovery & Planning (Tuần 1–3) 9](#_Toc208577809)

[Giai đoạn 2 — POC (Tuần 4–9) 10](#_Toc208577810)

[Giai đoạn 3 — Pilot (Tuần 10–17) 10](#_Toc208577811)

[Giai đoạn 4 — Rollout toàn công ty (Tuần 18–29) 10](#_Toc208577812)

[Giai đoạn 5 — O&M (Tuần 30 → liên tục) 11](#_Toc208577813)

# 1 — Tóm tắt chiến lược

Triển khai Ollama on-premises làm **AI engine nội bộ**; xây **RAG layer (vector DB)** để tăng chính xác trả lời; expose AI qua **AI Assistant microservice** tích hợp với API Gateway + IAM; chạy rollout theo 4 pha: Discovery → POC → Pilot → Rollout toàn công ty → Vận hành & tối ưu.

# 2 — Nhân sự & vai trò (gợi ý)

* Project Sponsor (TP/CTO)
* PM (1) — điều phối, timeline, stakeholder
* Lead Architect (1) — thiết kế kiến trúc, sizing
* Infra/DevOps (2–3) — deploy Ollama cluster, gpu/cpu nodes, storage, backups
* Backend Engineers (3–6) — tích hợp microservices, API, RAG middleware
* Data Engineer (1–2) — ETL văn bản, embeddings pipeline
* ML Engineer (1–2) — fine-tune, LoRA, prompt engineering
* Security Engineer (1) — IAM, encryption, audit logging
* QA/Tester (2) — test chức năng & security tests
* Business Analyst / Domain SMEs (HR, Legal, Finance) (2–4)
* Change/Training Lead (1–2) — đào tạo người dùng, adoption

# 3 — Giai đoạn & timeline ước lượng (tham khảo)

Tổng: ~5–6 tháng để rollout có kiểm soát, có thể rút ngắn/ kéo dài tùy tài nguyên.

1. Discovery & Planning — 2–3 tuần
2. POC (RAG + Ollama basic) — 4–6 tuần
3. Pilot (1–2 modules, ~200 users) — 6–8 tuần
4. Rollout mở rộng (toàn bộ hệ) — 8–12 tuần
5. Vận hành & tối ưu liên tục — ongoing

(Thời gian có thể song song: infra & ETL làm trước, dev microservices triển khai parallel)

# 4 — Giai đoạn 1: Discovery & Planning (2–3 tuần)

Mục tiêu: xác nhận scope, dữ liệu, yêu cầu bảo mật, sizing phê duyệt.

Tasks:

* Khảo sát dữ liệu: loại file, volume (GB/TB), tỷ lệ văn bản nhạy cảm.
* Chọn vector DB: Milvus / Weaviate / pgvector — đánh giá theo latency, backup, ops skill.
* Sizing: ước tính GPU/CPU, RAM, disk IOPS dựa trên QPS kỳ vọng.
* Quy định bảo mật & compliance: DLP, data retention, PII masking.
* Định nghĩa KPIs & SLOs (latency, availability, precision/recall).
* Lập risk register & mitigation plan.

Deliverables:

* Technical design doc (incl. network, firewall, VLAN, storage).
* Cost estimate & procurement list.
* Acceptance criteria cho POC.

Acceptance:

* Tài liệu hoàn tất, budget approved, hardware/cloud resource allocated.

# 5 — Giai đoạn 2: POC — RAG + Ollama (4–6 tuần)

Mục tiêu: Chứng minh khả năng RAG + Ollama trả lời tài liệu nội bộ với latency & privacy chấp nhận được.

Tasks:

1. Infra & Environment
   * Deploy Ollama runtime (single node) trên máy chủ on-prem (hoặc private cloud).
   * Triển khai vector DB thử nghiệm (Milvus/pgvector).
   * Kết nối storage cho dữ liệu văn bản (shared filesystem/S3-compatible).
2. Data pipeline (mini ETL)
   * Lấy sample dữ liệu: 1k–5k docs (PDF, DOCX, email).
   * OCR nếu cần, text extraction, cleanup, chunking (chunk size 500–1000 tokens).
   * Tạo embeddings (open-source encoder — hoặc model embedding trong Ollama nếu có).
   * Index vào vector DB.
3. RAG Middleware
   * Triển khai microservice “Retriever” trả về top-k docs + metadata (k = 3–5).
   * Thiết kế prompt template: system prompt + retrieved context + user query.
4. Integrate with Ollama
   * Gửi prompt + context tới Ollama API; nhận và hiển thị output.
5. Eval & Metrics
   * Tạo bộ test Q&A ~100 câu từ SMEs, đo accuracy, hallucination rate, latency.
6. Security baseline
   * SSL/TLS, internal network only, request-level audit logging.

Deliverables:

* Running demo: chatbot/RAG trả lời đúng với tài liệu sample.
* Report: accuracy, latency, resource usage.

Acceptance:

* Latency trung bình < 2s retrieval + < 3s generation (target tuỳ model).
* Accuracy ≥ threshold (do SME định nghĩa), hallucination rate giảm bằng citation links.

**Kết quả mong đợi:** POC chứng minh RAG + Ollama hữu dụng để tra cứu và tóm tắt văn bản.

# 6 — Giai đoạn 3: Pilot (6–8 tuần)

Mục tiêu: Triển khai cho 1–2 mô-đun (ví dụ: Quản lý văn bản + Trình ký điện tử) cho ~100–300 users.

Tasks:

1. Production-grade infra
   * Ollama cluster: scale out (replicas), load balancer, health check.
   * Set up persistent logs, metrics (Prometheus + Grafana), tracing (Jaeger).
2. Fine-tune / Prompt-engineering
   * Fine-tune nhỏ (LoRA) trên template văn bản nội bộ (policy, quyết định).
   * Xây prompt templates cho từng luồng nghiệp vụ.
3. Integration
   * Tích hợp AI Assistant microservice vào eOffice modules (API contract).
   * Implement RBAC: verify caller’s role via IAM before returning sensitive content.
4. UX
   * Thêm UI: suggestions, "cite source" links, confidence score, “view original doc”.
5. Monitoring & Ops
   * Setup alerting (latency, errors, high hallucination).
   * Implement rate-limits & caching (popular queries).
6. Training & Change Management
   * Tài liệu user guides, 2 workshops for pilot users.
7. Security & Audit
   * SIEM integration, request/response audit, redaction for PII (if needed).

Deliverables:

* Pilot-ready system, onboarding docs, training sessions, rollback plan.

Acceptance:

* Pilot users adoption ≥ X% (eg 50% weekly users) AND reduction TTR (time-to-response) cho nghiệp vụ ≥ 30%.
* No major security incidents; logs audited.

# 7 — Giai đoạn 4: Rollout toàn công ty (8–12 tuần)

Mục tiêu: Mở rộng ra tất cả module, tối ưu chi phí & performance.

Tasks:

* Scale infra (thêm nodes, shard vector DB nếu cần).
* Migrate more docs: full index pipeline, incremental indexing (change data capture).
* Implement multi-model strategy: model nhẹ cho simple chat (CPU quantized), model mạnh (GPU) cho generation/fine-tune.
* Implement governance: model-card cho từng model, data provenance.
* Automate deployment: IaC (Terraform/Ansible), CI/CD pipelines.
* Onboard all departments, run training + feedback loops.

Deliverables:

* Production system cho toàn công ty, runbook, SLOs signed-off.

Acceptance:

* System supports expected concurrency (QPS) with SLOs met.
* Business KPIs met (response time, reduced manual hours).

# 8 — Giai đoạn 5: Vận hành & Tối ưu (ongoing)

* Đặt schedule retrain/LoRA updates (monthly/quarterly).
* Monitor drift, update index, archiving policy.
* Feedback loop: collect false answers → retrain/adjust prompts.
* Cost optimization: quantization, mixed-precision, model distillation.

# 9 — Hạ tầng & sizing tham khảo

(ước lượng cho 1000+ users, concurrency peak 200–300 QPS — điều chỉnh thực tế)

Option small/medium/large:

* Small (poC / pilot): 1–2 GPU (NVIDIA A10 / A40) + 4 CPU nodes, 128GB RAM, 1–2 TB NVMe.
* Medium (production moderate): 2–4 GPU (A100/A30), 6–8 CPU nodes, 512GB RAM, 5–20 TB storage (S3/MinIO).
* Large (high throughput): GPU farm 8+, auto-scaling CPU nodes, dedicated vector DB cluster (replicas + shards), high IOPS storage, k8s cluster.

Network & Security:

* Internal-only endpoints, NAT egress locked down.
* TLS everywhere, mTLS for service-to-service.
* Key management (Vault), encrypted at rest.

# 10 — Bảo mật & Quyền riêng tư (must-have)

* On-premises deployment để bảo vệ dữ liệu.
* IAM integration: Keycloak / Azure AD SSO; RBAC checks in AI Assistant.
* Request-level access control: khi retriever trả metadata, filter theo permission.
* Audit logs: tất cả prompts/responses (cân nhắc masking PII).
* DLP: detect & redact PII from user inputs & outputs when required.
* Model usage policy: không cho model truy cập data ngoài scope; model-card & allowed use cases.

# 11 — RAG engineering (best practices)

* Chunk size: 500–1000 tokens, store metadata: doc id, section, timestamp.
* Use hybrid score: embedding similarity + BM25 for retrieval.
* Provide citations: attach top-k excerpts with source links to reduce hallucination.
* Keep “context window” limited; use selective retrieval + summarization.
* Freshness: incremental re-indexing on document updates.

# 12 — Testing & QA

* Unit + integration tests for retriever & generation flows.
* Security tests: pen-test internal, check exfiltration vectors.
* Evaluation dataset (SME-labeled) for ongoing QA; measure accuracy, F1, hallucination.
* Canary rollout for model updates.

# 13 — Rollback & Contingency

* Blue/green release for AI Assistant service.
* Versioned indices & model snapshots -> rollback to previous model quickly.
* Circuit breaker: if AI latency/error spike, fallback to rule-based responses or cached templates.
* Backup vector DB snapshots daily.

# 14 — KPIs & Success Metrics

* Business:
  + Time saved per request (avg) — target ≥30% reduction.
  + **of manual steps automated.**
* Technical:
  + Request latency (retrieve + generation) — p95 < 4s (target).
  + Availability SLO 99.9% (or agreed).
  + Hallucination rate < X% (business-defined).
  + Query throughput / CPU & GPU utilization.
* Adoption:
  + % active users (weekly/monthly).
  + NPS internal users for AI assistant.

# 15 — Chi phí & kinh phí (rất sơ bộ)

* HW đầu tư (GPU servers): tùy cấu hình — có thể vài chục tới 100k+ USD cho cluster.
* SW: vector DB OSS (Milvus/pgvector) miễn phí; enterprise support optional.
* Nhân lực: team 6–12 người trong giai đoạn 3 tháng đầu.  
  (Chi phí cụ thể cần RFP/quote sau sizing chính xác.)

# 16 — Ví dụ kỹ thuật nhanh (có thể dùng ngay)

## 16.1 — Ví dụ Modelfile (ví dụ tối giản)

# Modelfile (ví dụ cho Ollama)

from: registry/model/llama3:latest

name: eoffice-hr-assistant

system:

- |

You are EOffice Assistant. Use company policy documents and cite sources.

If unsure, answer "Tôi không chắc" and reference docs.

## 16.2 — Mẫu Prompt Template (RAG)

SYSTEM:

You are EOffice Assistant. Use only the documents provided in CONTEXT. Always cite source doc ids.

CONTEXT:

<<retrieved\_docs: [{"id":"doc-123","text":"...","meta":{"title":"Quy trinh nghi phep"}} , ...]]>

USER:

{user\_query}

INSTRUCTION:

- Provide concise answer (max 200 words).

- Provide source citations like [doc-123, p.3].

- If ambiguous, ask a clarifying question.

**16.3 — Mẫu gọi API (pseudo)**

POST http://ollama.internal:11434/v1/responses

Content-Type: application/json

Authorization: Bearer <service-token>

{

"model": "eoffice-hr-assistant",

"prompt": "<assembled prompt above>",

"max\_tokens": 512,

"temperature": 0.0

}

# 17 — Rủi ro chính & mitigation (ngắn gọn)

* Hallucination → mitigation: RAG + citations + conservative temperature + post-filtering.
* Leakage dữ liệu → on-prem + IAM + DLP + audit.
* Chi phí HW → bắt đầu bằng quantized CPU models cho phần không đòi hỏi generation nặng, sau đó scale GPU.
* Tốc độ → cache, use smaller models for simple Q/A, heavy gen on-demand.

# 18 — Acceptance checklist trước khi mở rộng

* POC pass accuracy & latency targets.
* IAM + audit logging hoạt động.
* Backup & rollback verified.
* Training completed + documentation sẵn sàng.
* On-call rota & runbook sẵn.

# 19 — Gantt Timeline (ước lượng ~6 tháng)

## Giai đoạn 1 — Discovery & Planning (Tuần 1–3)

| **Task** | **Duration** | **Dependencies** |
| --- | --- | --- |
| Thu thập yêu cầu, phân tích nghiệp vụ | 1w | - |
| Đánh giá dữ liệu (loại file, volume, nhạy cảm) | 1w | - |
| Chọn vector DB, sizing hạ tầng | 1w | Song song |
| Thiết kế kiến trúc high-level | 1w | Thu thập yêu cầu |
| Approval budget + resource | 0.5w | Thiết kế kiến trúc |

## Giai đoạn 2 — POC (Tuần 4–9)

| **Task** | **Duration** | **Dependencies** |
| --- | --- | --- |
| Deploy Ollama runtime (single node) | 1w | HW ready |
| Deploy vector DB (Milvus/pgvector) | 1w | - |
| Data pipeline: OCR, extract, chunk, embed | 2w | DB ready |
| RAG middleware (retriever API) | 1.5w | Vector DB |
| Prompt template & tích hợp Ollama API | 1w | Retriever |
| SME test set (100 Q&A), đo accuracy | 1w | Pipeline + retriever |
| Security baseline (TLS, audit logging) | 1w | Infra ready |
| Demo + báo cáo POC | 0.5w | All trên |

## Giai đoạn 3 — Pilot (Tuần 10–17)

| **Task** | **Duration** | **Dependencies** |
| --- | --- | --- |
| Triển khai Ollama cluster (multi-node, LB) | 1.5w | POC pass |
| Setup monitoring (Prometheus, Grafana, SIEM) | 1w | Cluster |
| Fine-tune LoRA domain-specific models | 2w | SME data |
| Tích hợp AI Assistant service vào eOffice (2 modules: Văn bản, Trình ký) | 3w | Retriever stable |
| RBAC/IAM integration (SSO/AD) | 2w | Gateway |
| UI/UX (chatbot widget, cite link, feedback button) | 2w | API ready |
| Training & onboarding ~200 users | 1w | Pilot infra |
| Pilot run & feedback loop | 2w | Training done |
| Báo cáo pilot (accuracy, adoption, latency) | 0.5w | Run complete |

## Giai đoạn 4 — Rollout toàn công ty (Tuần 18–29)

| **Task** | **Duration** | **Dependencies** |
| --- | --- | --- |
| Scale infra (GPU nodes, vector DB cluster) | 2w | Pilot success |
| Index toàn bộ tài liệu (bulk load + incremental update) | 3w | Data pipeline stable |
| Multi-model strategy (light vs heavy models) | 2w | Infra scale |
| CI/CD pipelines + IaC | 2w | Cluster |
| Tích hợp vào toàn bộ module eOffice (HR, Thanh toán, Phòng họp) | 4w | API stable |
| Onboarding toàn công ty (>1000 users) | 2w | Integration done |
| Training sessions (per department) | 2w | Onboarding |
| KPI tracking (usage, TTR, adoption) | 1w | Rollout done |

## Giai đoạn 5 — O&M (Tuần 30 → liên tục)

| **Task** | **Duration** | **Dependencies** |
| --- | --- | --- |
| Monitoring + alerting liên tục | Ongoing | Rollout |
| Feedback → retrain/fine-tune | Mỗi 2–4w | Adoption |
| Incremental indexing (new docs) | Ongoing | Data pipeline |
| Cost optimization (quantization, caching) | Ongoing | Usage patterns |
| Security audit & compliance check | 6 tháng/lần | Production |

**Trục thời gian (tóm tắt)**

* **Tháng 1:** Discovery (3w)
* **Tháng 2–3:** POC (6w)
* **Tháng 3–4:** Pilot (8w)
* **Tháng 5–6:** Rollout toàn công ty (12w)
* **Sau tháng 6:** O&M liên tục