FlologixAI Enterprise RAG Datasheet

SELF HOSTED -RAG - ORCHESTRATION

TURN YOUR PRIVATE LLM INTO A COMPANY WIDE KNOWLEDGE ENGINE: SECURE SEARCH, SUMMARIZATION, AND Q&A OVER YOUR DATA.

WHAT IS ENTERPRISE RAG?

- OBJECTIVE: DEPLOY AN ENTERPRISE GRADE RETRIEVAL AUGMENTED GENERATION (RAG) SYSTEM THAT KEEPS DATA PRIVATE AND IMPROVES DECISION MAKING ACROSS TEAMS
- SCOPE: : VECTOR DATABASE + AUTOMATED INGESTION + PRIVATE RETRIEVER AUGMENTED GENERATION + ADMIN & ANALYTICS
- OUTCOME: ACCURATE, GROUNDED ANSWERS WITH CITATIONS; FASTER KNOWLEDGE ACCESS; OPERATIONAL WORKFLOWS CONNECTED TO AI.

CORE CAPABILITIES

- CONNECTORS: FILE SHARES, SHAREPOINT/ONEDRIVE, GOOGLE DRIVE, EMAIL, S3/AZURE BLOB, DBS (READ ONLY).
- PRE PROCESSING: OCR, NORMALIZATION, DEDUPLICATION, QUALITY FILTERS, OPTIONAL PII REDACTION.
- INGESTION: CHUNKING STRATEGY, EMBEDDINGS, SCHEDULES, CHANGE DETECTION, VERSIONED JOBS.
- RETRIEVAL: TOP K SIMILARITY, HYBRID/METADATA FILTERS, OPTIONAL RE RANKING; SOURCE CITATIONS.
- GENERATION: GROUNDED ANSWERS WITH CONTEXT WINDOWS SIZED TO THE MODEL; CITATION LINKS.
- ADMIN: ROLE BASED ACCESS (RBAC), ANALYTICS DASHBOARDS, CONNECTOR STATUS & ERROR QUEUES.

WHAT YOU GET (DELIVERABLES)

- ARCHITECTURE & DEPLOYMENT OF RAG STACK (VECTOR DB, INGESTION, RETRIEVER, LLM INTERFACE).
- SOURCE DISCOVERY & INGESTION DESIGN; PRE PROCESSING & REDACTION PIPELINE.
- ADMIN CONSOLE (DASHBOARDS, LOGS, METRICS); USAGE & RELIABILITY ANALYTICS.
- EVALUATION PACK: RETRIEVAL & ANSWER QUALITY TESTS; ACCEPTANCE CRITERIA & RUNBOOKS.
- INTEGRATIONS: SLACK/TEAMS, TICKETING (E.G., JIRA), NOTIFICATIONS, REPO

PERFORMANCE AND SCALING

- SCALES WITH VLLM/GGUF HOSTING; CACHING FOR HOT PROMPTS; ASYNC BATCH EMBEDDINGS.
- SHARDED OR REPLICATED VECTOR DB; ANN INDEXES; STREAMING RESPONSES.
- SIZING GUIDANCE FOR GPUS/CPU

SUCCESS METRICS (KPIS)

- SEARCH TIME REDUCED (MIN/QUERY).
- ANSWER COVERAGE (% QUERIES WITH USEFUL CITATIONS).
- CASE DEFLECTION/FIRST TOUCH RESOLUTION (WHERE APPLICABLE).
- ADOPTION (ACTIVE USERS/WEEK) & RELIABILITY (P95 LATENCY, UPTIME)

SECURITY & COMPLIANCE

- DEPLOYED ON PREM OR PRIVATE CLOUD; CLIENT OWNS MODELS AND VECTOR DB..
- ENCRYPTION IN TRANSIT/AT REST; NETWORK ISOLATION, IP ALLOWLISTS; SSO/JWT; RBAC BY ROLE/TEAM.
- AUDIT LOGGING FOR QUERIES, ACCESS, APPROVALS; CONFIGURABLE RETENTION; EVIDENCE EXPORTS.
- POLICY BASED INGESTION ALLOW/DENY LISTS; HUMAN APPROVALS FOR SENSITIVE SOURCES (OPTIONAL).

EVALUATION & QUALITY

- RETRIEVAL: RECALL@K, PRECISION@K, MRR; COVERAGE & FRESHNESS CHECKS.
- ANSWER QUALITY: GROUNDEDNESS/FAITHFULNESS, COMPLETENESS; ERROR TAXONOMY.
- OPS: LATENCY BUDGETS (P95), UPTIME, THROUGHPUT, COST PER 1K QUERIES.
- FEEDBACK: HUMAN IN THE LOOP RATING LOOP; CONTINUOUS REGRESSION TESTING

TIMELINE & EFFORT (TYPICAL 4-8 WEEKS)

- 1) DISCOVERY & ARCHITECTURE
- 2) CONNECTORS & INGESTION PIPELINE
- 3) RETRIEVAL TUNING & EVALUATION
- 4) ADMIN DASHBOARDS, TRAINING, UAT
- 5) GO LIVE & HANDOVER.

TYPICAL STARTING PACKAGE

CONTACT US FOR A SCOPED QUOTE BASED ON DATA VOLUME AND MODEL SIZE.

FLOLOGIXAI ENTERPRISE READY PRIVATE AI WITH RAG AND ORCHESTRATION

FLOLOGIXAI.COM