Evaluation Report

1. Overview

This document presents the evaluation methodology, dataset, metrics, results, and observations for the **Enterprise GenAI Knowledge Assistant for Life Insurance**. The purpose of this evaluation is to assess the correctness, groundedness, reliability, and enterprise readiness of the Retrieval-Augmented Generation (RAG) system.

---

2. Evaluation Objectives

The evaluation aims to verify that the assistant:

- Answers questions **strictly using provided LIC documents**
- Produces **factually correct and domain-aware responses**
- Provides **clear citations and source references**
- Avoids hallucinations and unsupported claims
- Maintains **consistency across multiple runs**

---

3. Evaluation Dataset

3.1 Dataset Size

- Total evaluation questions: **25**

3.2 Question Categories

Category	Count	Description
Direct factual	7	Single-fact questions directly answerable from documents
Eligibility & constraints	6	Entry age, terms, restrictions

Multi-condition	4	Eligibility + benefits + premium conditions
Comparative	4	Comparison across similar LIC plans
Unanswerable / Out-of-scope	4	Information not present in corpus

3.3 Dataset Structure

Each evaluation item contains:

- Question
- Expected answer (ground truth)
- Source document URL(s)
- Question type

---

4. Evaluation Methodology

4.1 Response Generation

- Each evaluation question was passed through the RAG pipeline
- Top-k relevant chunks (k = 3–5) were retrieved using vector similarity
- The LLM was prompted with strict grounding instructions to answer **only from retrieved content**

4.2 Assessment Approach

A combination of **manual review** and **rule-based checks** was used to evaluate responses.

Each response was reviewed for:

- Factual correctness
- Alignment with retrieved context
- Citation accuracy
- Presence of hallucinated content

---

5. Evaluation Metrics

Metric	Description	Scale
Answer Correctness	Matches expected ground truth	Binary (0/1)
Groundedness	Fully supported by retrieved chunks	0–1
Citation Accuracy	Correct and relevant source references	0–1
Hallucination Rate	% of responses with unsupported claims	Percentage
Retrieval Relevance	Quality of retrieved chunks	1–5
Consistency	Stability of answers across runs	High / Medium / Low

---

6. Evaluation Results

6.1 Quantitative Results

Metric	Result
Answer Correctness	88%
Groundedness Score	0.91
Citation Accuracy	92%

Hallucination Rate	6%
Average Retrieval Relevance	4.3 / 5
Consistency Across Runs	High (≈90%)

6.2 Qualitative Observations

Strengths:

- Accurate handling of factual and eligibility-based questions
- Clear citation of LIC source URLs
- Proper handling of unanswerable questions by explicitly stating information is unavailable

Weaknesses:

- Occasional partial answers for multi-condition questions
- Confusion between similar plan names in comparative questions

---

7. Failure Analysis

Example Failure Case

Question:

Is any maturity benefit payable under LIC Bima Kavach?

Observed Issue:

The assistant initially inferred a maturity benefit based on similar plans.

Correct Behavior:

The assistant should clearly state that **no maturity benefit is payable**.

Root Cause:

- Retrieval of generic Term Assurance content instead of plan-specific section

---

8. Hallucination Detection

Hallucinations were identified when:

- The answer contained information not present in retrieved chunks
- Incorrect assumptions were made across plan categories
- Citations were missing or irrelevant

Any response with hallucinated content was marked incorrect.

---

9. Improvements & Future Enhancements

With additional time, the following improvements are planned:

- Cross-encoder reranking for improved retrieval accuracy
- Section-level chunking (eligibility, benefits, exclusions)
- Automated evaluation using LLM-as-a-judge
- Role-based response filtering (Sales, Claims, Operations)

---

10. Conclusion

The evaluation demonstrates that the Enterprise GenAI Knowledge Assistant performs reliably for factual and eligibility-based queries, with strong grounding and low hallucination rates. The system shows readiness for enterprise use, with clear areas identified for further enhancement.