

Generative AI based solution for Sales agent's productivity improvement in Life Insurance Industry

Sanjay Thawakar

Senior Vice President, AI Works & BPMA

Vibhu Srivastava

Vice President, AI Works

Divyan Kavdia

Vice President, AI Works

Abhishek Kumar

Senior Specialist, AI Works

Abhilash Pujar

Senior Manager, AI Works

Max Life Insurance

Abstract— In addition to overseeing the daily sales operations, life insurance (LI) companies employ a substantial number of proficient sales training and capability development teams. These teams are responsible for the ongoing recruitment and training of newly hired sales agents. With an accelerated growth in the Life Insurance Industry in India, the significance and criticality of these sales training teams have increased significantly. As Life insurance companies focus on rapid expansion plans, recruiting more insurance agent professionals and at the same time witnessing substantial attrition rate of the sales agents due to higher completion and number of opportunities in Life Insurance sales. These companies incur high costs of sales operations due to the limited availability of highly skilled sales training teams, hindering their ability to offer timely and targeted guidance and support to new sales team members beyond the initial classroom trainings. Consequently, this directly affects the company's sales performance, as a lower proportion of agents achieve sales productivity even after 6-12 months of joining the sales force.

In this paper, the authors present a solution to the above mentioned problem through A generative AI based solution- a conversational chat application driven by Generative AI models, which can serve as an intelligent virtual trainer for on-field sales agents, aiming to decrease the reliance on experienced human trainers and increase the productivity of new sales agents. The solution leverages a host of generative AI models, along with associated cutting edge transformer models (Instruction tuned LLMs, vector embedding models, speech transcription models), which were trained and contextualized for custom domain knowledge.

Keywords— Generative AI, Large Language Models, Lang chain, Embedding models, vector database and text retrieval, Insurance sales productivity management, Insurance sales training

1. INTRODUCTION

1.1. Background

The Life insurance industry in India witnesses accelerated growth with Life insurance company's premiums growing by

18% in FY23 as compared with FY22. In terms of the actual new business premiums, Life Insurance companies collected Rs. 3.71 lakh crore (US\$ 44.85 billion) as the first-year premium in FY23 as against Rs. 3.14 lakh crore (US\$ 37.96 billion) in FY22[1]. This industry growth is fueled by the overall growing financial literacy of customers along with a number of progressive and transformative initiatives by government, modified and conducive regulations increased industry partnerships, product innovations, vibrant distribution channels and latest digital disruptions witnessed by the industry. Leveraging the fast industry growth Max Life Insurance has rapid expansion plans of opening 100+ new offices in 100 new cities, in FY24. Currently Max Life Insurance has a field sales force of more than 1L agents and associates spanned across 250 offices, present in 140+ cities in India. Life Insurance is a distribution driven market where till date majority of sales happen via advisors. These sales advisors are responsible for selling max Life insurance policies through various sales channels including Max Life core agency offices and with multiple partner banks. All the Max Life sales agents and associates are continuously trained by the Max Life Distribution capability development team.

1.2. Problem Statement

With a rapid business footprint expansion strategy, in the business environment of stringent competition, one of the biggest challenge for Max Life is to continuously maintain and enhance the capability and productivity of such a large sales force. Apart from setting up the new expansion office infrastructure, this business expansion involves, Agent Hiring, Agent Training, On-job productivity measurement & enhancement and agile customer support. This involves taking on multiple challenges such as –

- High attrition levels of this field force ~ 50% attrition, increasing the inflow of fresh and new sales associates to contribute to 40% of the sales force every year
- Difficult to train new associates in a short time owing to complexity of insurance products along

with regulations & monitoring around fair sales practices by regulatory authorities

- Hard to find and expensive to hire large number of experienced Life insurance sales trainers, making the trainer to trainee ratio very low (1 trainer for every 50 new agent hired) poorly impacting the quality of training and making it difficult for trainers to provide tailored support beyond initial 3 weeks of classroom training
- Continuous training requirements even for the vintage sales associates with rapid pace of new Life Insurance products being introduced in the market and requirement of staying up to date on financial markets/current affairs which may impact market linked insurance products outcomes and risk

The sales training process includes the sales capability development of sales agents right from the time of onboarding to continuous skill enhancements during the entire tenure with Max Life. It includes training the sales agents across life insurance sales cycle, includes trainings on end to end sales process, max life and competitor product offerings, handling customer communication and objection, and, is driven by experienced Life Insurance sales trainers. The entire sales and distribution training operations involves significant efforts and cost for continuous training and upskilling of sales force team members.

Given the above business scenario, the key problem statement was to ensure anytime anywhere availability of sales assistance to the field sales force (basis their personas), to ensure higher sales productivity increasing the revenue for the company and at the same time reducing the dependency on experienced human trainers thereby reducing the overall training costs.

1.3. Purpose of Research

The purpose of the research was to build an end to end generative AI (Large Language model) based conversational chat application, which can act as an intelligent virtual trainer and sales assistant to the on-field sales agents reducing the dependency of on ground sales agents on skilled trainers and experienced sales professionals. This chat application acts as a question answering bot interacting in natural, human-like conversations with the right context, tone, and intent of the query in multiple languages, to provide answers to all the sales and product related questions to every sales agent on the mobile anytime, anywhere.

2. SOURCE OF DATA

With the objective to build a chat application with specific Life insurance sales and Max Life products knowledge, the data used in the study is majorly proprietary to Max Life Insurance, except for data from a few financial aggregator websites for the industry and competitor product comparisons. Multiple workshops were held between the analytics team & the sales training team to chart a roadmap of the training journey for the newly joined sellers. Once the roadmap was charted, the next logical objective was to assess

and collate the relevant material to build a robust domain knowledge base.

The data was majorly segregated into 3 broad categories –

- A) *Product information data*- Life insurance products being offered, including information on their main features and benefits, coverage, exclusions, premium payments, calculation methods, claims process, additional available options, regulatory disclosures, and more
- B) *Sales Process training data* – Documents primarily focusing on the implementation of effective sales techniques, such as prospecting, lead generation, cold calling, and networking. Furthermore, they provide strategies for overcoming objections, handling rejections, and successfully closing sales. To ensure a comprehensive understanding of the sales process, the training team also facilitates various role-playing scenarios using scripts and templates, as well as presents case studies and success stories for analysis and discussion.
- C) *Industry, competition and financial markets data*– The industry and financial markets datasets primarily consisted of current affairs of financial markets, insurance industry related developments and competition product comparisons and information. These datasets were obtained through insurance aggregator websites and financial analyst datasets such as policybazar.com, Morningstar.com, after converting the website datasets into PDF formats.

Since most of the training material used by the Sales training team was originally intended for classroom training, the knowledge artefacts are not created linearly or as plain text – most of the data was available in the format of PowerPoint presentations and PDFs consisting of infographics, animations, nested tables, illustrations, audios and videos etc.

3. METHODOLOGY

The overall methodology involved applying the research to build a solution which can be used to solve the business problem at hand. The technical solution development involved a 6 step process, apart from gaining the functional business understanding of sales training and sales process, as first step. The solution development involved a study and curation of 250+ documents in the form of PowerPoint presentations, PDF, videos along with the content across Max Life website and various other financial aggregator websites as mentioned in the data description section.

FIGURE 1: 6-STEP PROCESS FOR BUILDING THE SOLUTION



A) Document Curation (Data extraction & Processing)-

As most of the datasets were in the form of unstructured PowerPoint presentations, PDF product brochures, videos and web-scraped text content, a number of document parsing and speech transcription models were used to

convert the data (PDFs, PowerPoint presentations, audios, videos, website HTML) into standardized linear 1000+ pages of PDF document with appropriate document structure to enable the LLM (Large Language models understand this data. Open AI's Whisper-1 transcription model was used for transcription of the video and audio datasets.

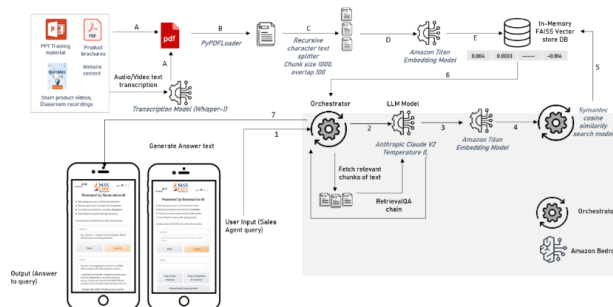
B) Creating document vector database

At the next step, the complete set of documents were broken down **into linear chunks of text (tokens)**- smaller sentences/ text with defined chunk size and overlap/split parameters to produce consistent sized chunks which is then converted into the **numeric 2 dimensional vector text embedding** [2] to form the part of complete knowledge database stored in the form of numeric dataset. A **Recursive Character Text Splitter** [3] was used to split the text into the required chunks, ensuring that Amazon's **Titan-Embedding LLM model** [2] is used to create the text embedding and these are then permanently stored as vectors in a PostGres database, to enable the LLM retrieve the appropriate answer to input seller questions

C) Solution architecture using the LLM models

Following the technical objective of the solution, an iterative LLM model framework development process was followed to leverage the generative AI Large Language models (Instruction tuned LLM) to enable generate appropriate **answers** and insights to seller questions. A number of different pre-trained LLM models such as gpt-3.5-turbo, Llama 1-13B-chat, Falcon 13B-chat, H20 GPT and Anthropic's Claude-V2 (AWS bedrock version customized for chat) were tried and evaluated. Basis the model's ability to understand user queries and generate factually correct and detailed responses, Anthropic's Claude models were selected for further development. Leveraging the key data components and the LLM framework, an orchestrator based architecture was build combining the key components of the solution by Andreessen Horowitz [4]

FIGURE-2 : SOLUTION ARCHITECTURE USING LANGCHAIN



The open-source orchestration, Langchain [6], framework was used as it can be used as a simple API, has a distributed architecture, and a variety of features for managing LLMs, such as load balancing, fault tolerance, and security.

The **anthropic' s claude-v2 LLM model was used** (hyper parameters-temperature 0, maximum tokens 512) to get most precise and factual answers avoiding any hallucinations and limit the answer response to a maximum of ~300 words) with the **Conversational Retrieval Chain (Lang chain framework) which** enables to translate the sales query and fetch relevant chunks of text the retrieved answer by running the **semantic** similarity search model (**using cosine similarity model**) on the previously stored vector database.

The FAISS (Facebook AI Similarity Search) library was used for efficient similarity search and clustering of dense vectors. The Prompt template used, was defined as the persona as helpful Max Life insurance agent with a polite tone restricting the answers to only Financial and Life insurance domain related questions

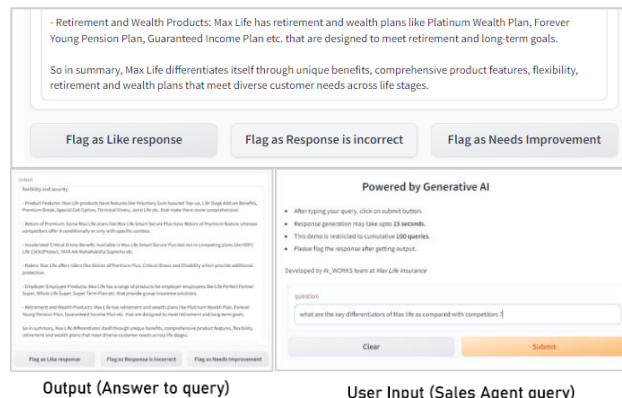
The research solution was build using python language and hosted on AWS private compute instances (EC2).

A **simple user interface** using the Gradio UI [6] interface was developed for the users (sales agents) to ask sales related questions and get the relevant answers. The UI also featured Human feedback loop where the user can mark if the answer was accurate. This feedback data collected was then used to further improve the quality of responses.

4. RESULTS

As multiple models were tried for both Embedding and

FIGURE-3: USER INTERFACE



LLMs (mentioned in the methodology section), the model outputs were validated through human feedback loop enabled by actual business users. Based on the test sample of ~1000 responses, the selected LLM model accuracy of Claude V2 model was established to be 92% as compared with the GPT 3.5 Turbo model outcomes. While other techniques such as automatic model evaluation (feeding answers as questions to LLM) were used to test the efficacy, actual user feedback was considered to be the final evaluation criteria as being most robust.

TABLE 1: COMPARING LLM – CLAUDE & GPT 3.5-TURBO

Question Category	# Questions	Claude V2		GPT 3.5 Turbo	
		Correct Answers as expected	% Accuracy	Correct Answers as Expected	% Accuracy
Sales and Pitching	216	210	97%	210	97%
Products Related (Recommendations and variant information)	150	108	72%	91	60%
Competition comparisons	126	114	90%	54	43%
Financial knowledge	270	264	98%	264	98%
Sales pitching, Conversation Starters and Objection Handling	174	168	97%	168	97%
Total	936	864	92%	787	84%

One of the major areas where the LLMs failed to answer was from complex tabular data having nested tables, merged cells and column heading. This made it extremely difficult to get the correct answers for questions related to competitor product comparisons or inferring the underwriting grids etc. Hence such tables present in the sales training documents were transformed in the form of multiple transposed single tables for the LLMs to infer the data and provide the correct answers.

FIGURE 4: RAW TABLE TRANSFORMATION

Nested Table					Transformed Table				
Variants	Age 30-45				Age	Investment	Product	Option	Outcome
	Investment SL		Investment >SL		30-45	SL	A	A	Z
	Product A	Product B	Product A	Product B					
	Option A	2	4	5	4				
Option B	4	5	2	3	30-45	>SL	B	C	4
Option C	5	5	4	4					

Similarly, iterative prompt engineering approaches such as Passing the current date time to get the most recently updated information along with explicit step by step calculation methodologies were fed in to improve answer accuracy.

Further, multiple vector search methodologies such as cosine similarity, MMR (maximum marginal search) [7] etc., were experimented with. During the evaluation phase it was established that MMR did not provide enough lift in accuracy along with requiring higher compute time for the given specific use case.

Another experiment was done leveraging the Query expansion [8] technique was used and its application was evaluated based on the accuracy and quality of answers. The Query expansion technique leverages the LLM to split the user question into maximum 3 sub-questions which are independent from each other and combines the answers of these 3 questions after the answer retrieval process. While, this experiment increased accuracy with long and complex questions, but generated larger un-necessary text when the questions were very specific such as “what is Max life insurance claims paid ratio?”. Hence this experiment dint yields high incremental efficacy while increasing response time.

5. CONCLUSION

While Generative AI is still evolving and is in discovery phase for most business applications. The solution developed with the defined LLM architecture and selected components was established to be the most accurate (92% accuracy) architecture along with the selected for the given use case and problem statement.

The research based solution comprehensively covers end to end support for the new and vintage sales force agents across all kinds of questions and suggestions such as-

- Sales Pitch Assistance (ready pitches and products)
- Product recommendations based customer needs and persona along with Give-get scenarios
- In-depth MLI product information and comparison of various product variants
- Compare competitor products (across features, risk-reward, pricing) with Max Life products
- Objection Handling
- Awareness of financial instruments and current affairs related to financial markets
- Policy issuance underwriting guidelines (Non-medical limits, policy premium calculations etc.)

The solution also supports to answer the questions in multiple vernacular languages. Apart from the estimated direct business impact, the solution also enables significant qualitative benefits driven by improving the quality of trainings with higher physical trainer’s availability, enabling High quality, smarter customer interactions which will help improve Max Life brand along with an impact on sales conversions. It also helps in standardization of sales practices (answering within the defined framework) driving sales process consistency, unified product knowledge and clear products understanding to reduce chances of miss-selling, improving the quality of sale and increasing customer delight.

REFERENCES

- [1] <https://www.ibef.org/industry/insurance-sector-india>
- [2] “Embeddings - Amazon Bedrock.” <https://docs.aws.amazon.com/bedrock/latest/userguide/embeddings.html>
- [3] “Recursively split by character |  Langchain.” https://python.langchain.com/docs/modules/data_connection/document_transformers/text_splitters/recursive_text_splitter
- [4] Rajko Radovanovic, Matt Bornstein. “Emerging Architectures for LLM Applications.” Andreessen Horowitz, 20 June 2023, a16z.com/emerging-architectures-for-llm-applications/.
- [5] “Introduction |  Langchain.” *Python.langchain.com*, python.langchain.com/docs/get_started/introduction. Accessed 28 Sept. 2023.
- [6] “Gradio,” <https://www.gradio.app/docs/interface>.
- [7] “Select by maximal marginal relevance (MMR) |  Langchain.” https://python.langchain.com/docs/modules/model_io/prompts/example_selectors/mmr
- [8] R. Jagerman, “Query expansion by prompting large language models,” *arXiv.org*, May 05, 2023. <https://arxiv.org/abs/2305.0365>