J. Electrical Systems 20-2 (2024): 2568-2581

[1]Manoj Kumar Manmathan
[1]Dr. Pankaj Agarwal
[2]Suraj Ravi Shiwal
[1]Nitin Bhore
[1]Shagun Singal
[1]Bhaskar Saha

Organization-Wide Continuous Learning (OWCL): Personalized AI Chatbots for Effective Post-Training Knowledge Retention

Journal of Electrical Systems

Abstract: - This research paper proposes the Organization-wide Continuous Learning (OWCL) system, employing AI chatbots to enhance post-training follow-up activities for employee knowledge retention. Integrating various functionalities (for continuous learning) and leveraging very large language models (vLLM), OWCL provides personalized learning through spaced repetition of topics, adaptive learning, gamification, etc. Here, we discuss a prototype built on the Gemini API, which demonstrated impressive accuracy (over 94%) in core functionalities like question generation and answer evaluation, showcasing the potential of vLLM to revolutionize post-training activities viz recall, revision, and application. With an overall accuracy of 85%, OWCL presents a balanced and practical approach, harnessing cutting-edge technologies while remaining accessible, resource-efficient, and reasonably fast, ensuring both cost-effectiveness and swift implementation.

Keywords: AI, LLM(Large Language Model), Microsoft Phi2, Google Gemini API, Organizational Post-Training Follow-up, Continuous Learning, Knowledge retention, Forgetting Curve, Spaced repetition.

I. INTRODUCTION

The effective transfer of knowledge from training programs to workplace application remains a persistent challenge across various industries. Our experience, corroborated by a recent survey conducted with employees across different organizations, highlights a common issue: employees often struggle to retain newly acquired knowledge if it is not regularly used in their daily tasks or revisited through revision. This research project centers on this critical theme, investigating the potential of Artificial Intelligence (AI) to mitigate knowledge retention difficulties and enhance the overall effectiveness of training programs.

A. Research Gap:

Post Training Follow-ups in Organizations + Continuous Learning Knowledge Base + Personalized Spaced Recall and Revision + Basic Prototype Implementation of a Very Large Language Model API-based chatbot for this use case + Evaluation of the prototype. We integrate multiple technologies to solve this unique use case. While AI-based solutions exist in the education field, solutions also exist for adaptive learning. Also, solutions exist in the industry to create a centralized knowledge base. However, not much research has been done on supporting automatic revision/recall for employees based on this large collection of information.

B. Target Audience:

Technology team, Learning & Development (including Certification training), HR etc. The project can be ported to any domain but will need more research.

C. Background 1:

Millions are invested in employee training programs each year around the globe, but a crucial aspect often remains unaddressed: the long-term retention of the acquired skills and knowledge.

Problem 1:

While the aim is to foster growth, innovation, and improved performance, across various domains, however the well-established psychological phenomenon known as the Forgetting Curve [1], reveals that without proper recall

[1]*Corresponding author: BITS Pilani, Dept of Computer Science & Information Systems, India. Manoj75.Code@gmail.com

[2]Prodigitalworx.inc, Data Science, Bangalore, India.

& revision of the information, a significant portion of the acquired knowledge from the training is susceptible to being forgotten within a few weeks to months' time. The absence of such post-training follow-up activities contributes to a significant reduction in employees' knowledge retention. This, in turn, diminishes the return on investment in employee development.

Example 1: Imagine attending a comprehensive training session but weeks later struggling to recall key concepts if not exposed to them in daily work.

Figure 1 Knowledge Retention of Employees belonging to diverse organizations, after initial Training, if not revised.

Figure 1 How often does organizations conduct Post-Training Follow-ups

Figures 1 and 2 present the results of a survey conducted among peer mates pursuing work-integrated MTech online, India, most of whom are employed in various private IT-based industries in India and some from abroad organizations. Eighty-five people responded.

Figure 1: 83% of respondents stated that they had forgotten most of the concepts/ terminologies within 1 week of the training. 13% people could retain the training information up to one month. 4% said they use the training information in daily use; hence they don't forget.

Figure 2: Among the respondents, 79% indicated that no post-training activities are conducted in their organizations, while approx. 15% reported rare occurrences of such activities. Additionally, up to approx. 6% of employees stated that their organizations conduct active follow-ups every month.

So, we deduce that most of the organizations (90% +) don't conduct active post-training revision/ recall activities. Furthermore, upon discussing this issue with two Technical Managers from a prominent Telecom company, they expressed that implementing an OWCL system would provide significant value to their employees, facilitating both information retrieval and active revision.

D. Background 2

Besides training documents, organizations amass extensive information in diverse formats such as manuals, presentations, training video transcripts, and files stored on network drives. This repository continually expands with additions like official meeting transcripts, emails, and daily work reports. Together, these resources constitute the Continuous Learning Knowledge Base (CLKB). However, employees typically engage with this unstructured information reactively, addressing specific needs or issues. Consequently, accessing and utilizing the CKLB can prove time-consuming and labor-intensive, particularly when data is scattered across various sources. Although top-tier and IT-related organizations commonly maintain centralized knowledge bases, many others encounter challenges in effectively consolidating and organizing their knowledge assets.

Problem 2: Hence, the challenge lies not only in retaining knowledge from training but also in facilitating efficient access and recall of the Continuous Learning Knowledge Base (CLKB) to support ongoing learning and problem-solving initiatives.

E. The larger Solution

The AI chatbot, leveraging various APIs or protocols, connects to the training documents stored in the organization's drive and the continuously expanding Continuous Learning Knowledge Base (CLKB). This chatbot autonomously generates context-based questions from the CLKB and offers prompts and support to employees for personalized revision and recall, all without the need for Learning and Development (L&D) intervention. Employing adaptive learning techniques, the chatbot dynamically adjusts the questions based on the employee's proficiency, thereby optimizing the learning experience.

F. The objectives of our research paper:

a. Explain the benefits of the OWCL project, as discussed above.

b. Take up a core functionality and implement it.

c. Evaluate & discuss the out-of-the-box performance of the new very Large Language Model used – Gemini API.

d. Target a reasonable performance (>80%). e) Report of a small language model, if used in place of Gemini API. While striving for an ideal AI system might be tempting, the reality is that such a pursuit would be cost-prohibitive and prone to fast obsolescence, given the rapid evolution of technology. We therefore propose a scalable approach that adapts alongside technological advancements, **starting with a feasible implementation and growing in complexity** as the AI industry progresses.

G. Model selection:

Recently, significant advancements in Large Language Models (LLM) research have led to models with improved capabilities and performance, and several companies are now offering public access to their LLMs through APIs and cloud services. While the locally installed, CPU-based, small-sized LLM (example: Microsoft Phi2) performance cannot be compared to a very large language model like Gemini, that could still be a good starting point. As we set the target for reasonable performance with fewer resources, the Gemini API far exceeds our expectations, and we believe other LLMs offer similar accuracies.

H. Organization-wide Continuous Learning (OWCL)

Our research endeavors to address the challenges in post-training follow-up activities for employee knowledge retention. We propose the concept of OWCL, a holistic approach aimed at enhancing knowledge retention through objectives such as personalized recall, revision, query assistance, employee assessment, and feedback. This approach contributes to the improvement of employees' skill sets, subsequently enhancing their efficiency. Additionally, we emphasize employee engagement through automated mechanisms like healthy peer competitions, gamification, and continuous access to relevant information throughout their careers, speech to text etc.

The OWCL strives to meet the above workforce demands through the utilization of cutting-edge technology, specifically AI chatbots. Leveraging data-driven insights, the OWCL identifies areas for improvement, personalizes revision learning content, and incorporates effective feedback and guidance mechanisms. This comprehensive approach ensures a tailored and efficient learning experience for employees.

I. Open-ended responses vs MCQs

The system extends beyond conventional Multiple-Choice Questions (MCQ) by incorporating open-ended responses. This integration enables a thorough assessment of knowledge through Natural Language Understanding (NLU), providing valuable insights into the extent of information retention by employees.

At the core of our system is an OWCL bot that integrates Large Language Models (LLMs) and various other modules. Our innovative strategy incorporates LLMs for multiple functions, including question-answer generation (QAG), user response evaluation, feedback, and chat with pdf (RAG) services. These are performed on the organizational documents and knowledge bases.

J. The spaced repetition system (SRS)

SRS is responsible for the customized spaced repetition of questions at intervals of immediate, 24 hours, 1 week, 1 month, and 3 months. The bot poses open-ended questions, evaluating employee responses almost instantly and displaying grades, feedback, context, and correct answers, thereby supporting the recall and revision. Employees can pose training-related questions, and the bot answers using RAG LLM. All interactions are saved for further integration into the spaced repetition system. This allows the bot to dynamically revise and evaluate employee learning, tailoring the approach based on past performance, automatically identifying knowledge gaps within individuals and teams, and automatically sending recommendations to the immediate bosses or training department. Functioning as a post-training Learning Assistant, the system can be activated for targeted training refresher, thereby maximizing knowledge retention.

K. Scope

Historically, training documents referred mainly to PDFs or other textual formats. Our initial prototype focus is on constructing a text-based system. Additionally, there's the capability for integration of APIs from diverse organizational channels like meetings, databases, work reports, training recordings, and emails to enhance continuous learning. This necessitates tailored preprocessing based on the specific data types involved.

L. Uniqueness

The implementation of LLM-based AI chatbots to address the challenges of post-training requirements in the organization, along with the continuous learning knowledge base, utilizing LLMs, represents a novel approach, currently not implemented globally. With limited existing studies, this research explores uncharted territory, shedding light on the distinctive features that set this technology apart. While organizations traditionally rely on MCQ for assessments, recent technological advancements allow for the integration of short-answer assessments, expanding the evaluation methodology at a reduced compute resource requirement.

The system facilitates a targeted approach to employee training by addressing specific topics guided by insights from the evaluation tool derived through historical data analysis, with a focus on areas of weakness. Its adaptive nature extends to the frequency of knowledge-refreshing questions, automatically adjusting to individual memory retention abilities and ensuring optimal and personalized learning experiences for each employee. Furthermore, the system exhibits seamless portability to various domains and adaptability with some retraining, showcasing its scalability and versatility in alignment with the company's expansion plans.

M. Benefits

The implications of this technology reach beyond the realm of novelty. Code & Text-based LLMs are now popular, stable, and affordable; hence, they can be implemented in any organization. By ensuring prolonged knowledge retention among employees, the OWCL bot keeps their skills at their peak, ultimately augmenting their contributions to the organization. Furthermore, the return on investment (ROI) is expected to be manifested through heightened productivity and efficiency, with reduced costs associated with repetitive training.

II. RELATED WORK

A. Spaced Recall & revision for knowledge retention

[1] Hermann Ebbinghaus was a German psychologist who founded the experimental psychology of memory. The "forgetting curve"—the loss of learned information—is sometimes referred to as the "Ebbinghaus Forgetting Curve.

[2] 'Rate of forgetting is independent from initial degree of learning across different age groups' explores the relationship between the amount of information learned and how well it is remembered. The Researchers compared memory in younger and older adults. They found that the rate of forgetting was the same for both groups. This means that forgetting happens at the same speed regardless of how much information is learned.

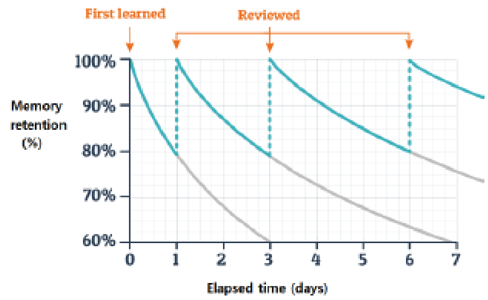

Figure 1. Ebbinghaus' forgetting curve and review cycle.

Figure 1 Courtesy: [15] shows that generally 80% is retained after 1hr and only 60% after 3 days. Review is mandatory to retain the learnings.

[3] 'Spaced Effect Learning' talks about spaced repetition—involves revisiting studied content at multiple, specifically selected time intervals to reinforce learning and facilitate long-term retention. Using question-based repetition, as opposed to passive reading/listening modalities, can help optimize this process. It mentions that Spaced learning has been used for training in multiple fields, including finance, management, technology, etc.

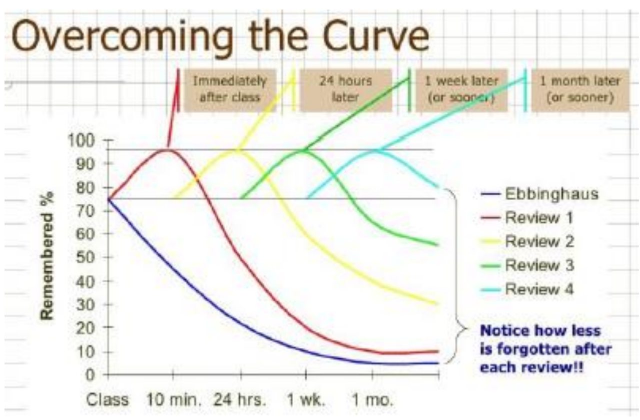

Figure 2 Courtesy: [16] also shows how the forgetting curve can be overcome by spaced repetitions at increasing intervals.

B. Importance of Post-Training Follow-up Activities for the Employees:

[4] 'Improving Training Impact Through Effective Follow-Up: Techniques and Their Application' emphasizes the importance of follow-up activities because organizational training alone is insufficient. It explores several

low-cost methods employers can use to support training and ensure skills and knowledge are applied in the workplace. These methods rely on peer and supervisor support to create a positive work environment.

[5] ‘Transfer of Training: The Known and the Unknown’ explores the gap between the increasing investment in training and the limited evidence of individual improvements. It focuses on generalization (applying learnings) and retention (maintaining learnings) over time.

[6] ‘Impact of Knowledge Sharing and Knowledge Retention on Employees Development’ finds that Knowledge Sharing & Knowledge Retention has a significant positive relation with employee development and sustainable competitive advantage mediates significantly between them.

C. LLMs:

[7] ‘Gemini: A Family of Highly Capable Multimodal Models’ discusses the features of Gemini by Google. Gemini Ultra model advances the state of the art in 30 of 32 of these benchmarks — notably being the first model to achieve human-expert performance on the well-studied exam benchmark MMLU and improving the state of the art in every one of the 20 multimodal benchmarks.

[8] ‘Microsoft Research blog’ mentions that Phi2 2.7B model surpasses some other 7B, 13B, 70B models.

D. Chatbots supported Learnings:

[9] The paper "A Question Answering and Quiz Generation Chatbot for Education" introduces a chatbot system for education, focusing on answer extraction and question generation in subjects like Social Studies and Science. It highlights the increasing popularity of chatbots facilitated by instant messaging services. The paper emphasizes the impact of NLP advancements on the growth of intelligent tutoring systems and references studies affirming that chatbots enhance student engagement in studies.

E. LLM-based QAG, answer evaluation, RAG

[10] studies about improving the quality of QA generated by LLMs even under the constraints of 1) a black-box (non-modifiable) question generation model and 2) lack of access to human-annotated references — both of which are realistic limitations for real-world deployment of LLMs.

[11] is research on the importance of context verification after the LLM-based question generation process, which helps remove hallucinating QA. Two experiments to automatically validate generated answers against a corpus are proposed.

[12] ‘Generating multiple choice questions from a textbook: LLMs match human performance on most metrics’ – This research indicates that the fine-tuned LLMs could generate questions competitive with human-authored questions.

[13] ‘A large language model-assisted education tool to provide feedback on open-ended responses’ - presents a tool that uses LLMs, guided by instructor-defined criteria, to automate responses to open-ended questions, delivering rapid, personalized feedback, enabling students to quickly test their knowledge and identify areas for improvement.

[14] ‘Retrieval-Augmented Generation for Knowledge-Intensive NLP Task’ acknowledges LLM’s ability to store factual knowledge in their parameters and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. They found that for language generation tasks, RAG models generate more specific, diverse, and factual language than a state-of-the-art parametric-only seq2seq baseline.

F. The findings from the above studies could be summarized as:

a. Employees forget most of the training learnings within a month’s time if they are not reviewed.
b. Irrespective of age, re-reading is not as effective as ‘recall and revision’ at spaced intervals. Generally, the suggested spacings for recall/revision are 10mins, 1hr, 24hr, 1week, 1month, 3months.
c. Organizations are spending a lot on employee training but are not sure how to measure effectiveness.

d. If the employees find it interesting, they will engage in continuous learning.
e. Chatbots are revolutionizing the world in multiple domains and keeping people engaged in learning.
f. Recent advancements in LLMs make them suitable for question-answer pair generation, user answer grading/evaluation, chat with pdf using RAG, etc.
g. Currently, the quality of the evaluation of user answers by the LLMs almost competes with that of humans, but there can still be mistakes.

III. APPROACH

The system is activated once the training is concluded.

A. System Design:

A modular system consisting of state-of-the-art technologies and several components:

B. Pre-processing, QAG, feedback & RAG. Data Ingestion & Preprocessing:

Facilitates importing various training materials (initially PDFs – these being the most important in trainings, later expanding to include recordings, images, codes etc.). For this prototype, manually selected a few pages from Azure training pdf. The system then chunks them - chunk size we chose 10000 tokens for Gemini QAG, for RAG 1200 as it depends on the embeddings as well. Overlap of 400 tokens.
Question-Answer Generation: The application iterates through contexts C, prompting the LLM to generate questions Q from C.
Each Q, C pair is again passed to the LLM and prompted to answer each question based on the given context. Thus, Irrelevant questions are identified during the answer generation process, allowing the system to skip them and proceed to the next question.

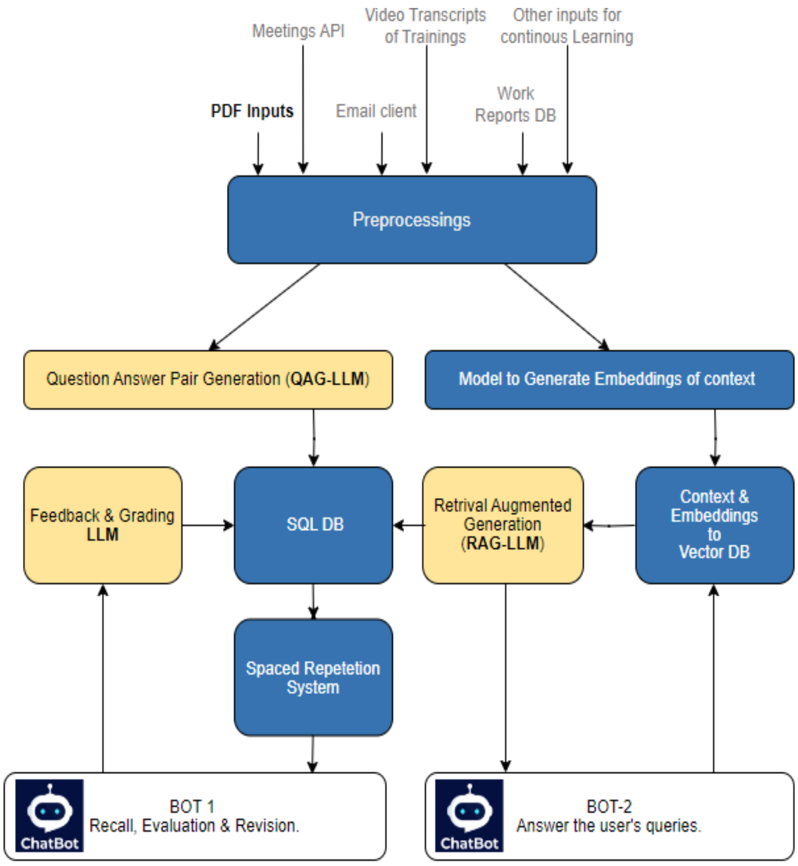

Figure 3 OWCL Assistant - Block diagram

C. BOT1, Grading & Feedback:

LLM is prompted to evaluate user responses to BOT-1 questions. LLM provides feedback, correct answers, and assigns grades. Bot1 acts as a practice tool, asking users questions on the user selected topics, repetition of questions is based on the SRS algorithm. It shows the feedback, grade, correct answer, context to the user, helps in recall and revision.

D. Database Management

Stores everything - generated questions, context, grades & feedback, user interactions etc.

E. BOT-2, User Inquiry and RAG Support:

As the Phi-2 did not perform well on QAG, we stuck to Gemini for the RAG implementation. Employees can request further explanation by asking questions on pdf. The LLM utilizes RAG (Retrieval-Augmented Generation) to retrieve relevant information from previously ingested training materials and provide comprehensive explanations or summaries. The system calculates the embedding vector for the user's question. Then, calculate the cosine distance between the user's question embedding and all stored chunk embedding vectors in the chromadb collection. It retrieves the "TOP_K" document objects with embedding vectors closest to the user's query. Concatenate them & feed them and the user's questions to the Language Model (LLM). Prompt the LLM again with the context and question to retrieve the relevant piece of information.

F. Spaced Repetition System (SRS):

The rule-based algorithm determines the optimal timing and sequence for presenting questions to users, considering individual performance history and course completion dates. The mathematical formula prioritizes questions based on specific criteria. Let Q represent the set of all questions, and Qasked denote the questions already presented to the user.

Next Question = argmax$_{q \in Q}$

$$\left[\alpha \cdot 1_{q \notin Q_{\text{asked}}} + \beta \cdot 1_{q \notin Q_{\text{24hrs}}} + \gamma \cdot 1_{q \notin Q_{\text{week}}} + \delta \cdot 1_{q \notin Q_{\text{month}}} + \epsilon \cdot 1_{q \notin Q_{\text{3months}}}\right]$$

In the above equation, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, $\boldsymbol{\epsilon}$ are weighting factors, and $\mathbf{1}_{\mathbf{condition}}$ is an indicator function evaluating to 1 if the condition is true and 0 otherwise. The criteria include prioritizing questions not yet asked and those not presented in the last 24 hours, last week, last month, and last three months, with respective weights adjusted to reflect the specified priority order. This formula ensures an optimal selection strategy, considering the user's historical interactions with the questions.

G. Implementation:

The prototype was developed in Python & UI with Gradio. We experimented with both Microsoft Phi2 model (small language model which works on CPU) & also the Google Gemini API. Our motto was to verify whether the LLMs could be used reasonably well, for this use case, for text-based PDFs. Implemented Prompt engineering for all three applications viz QAG, user answer evaluation & feedback, and RAG.

H. System Operation

a. Content Ingestion: The training department or the employees would upload their Training materials into the system for preprocessing.

b. QA Generation: The LLM analyzes the preprocessed content and generates a pool of personalized QA pairs.

c. Spaced Repetition: The SRS schedules revision sessions, selecting questions based on difficulty, previous performance, and the forgetting curve.

d. Interactive Revision: The system presents the employee with a question from the database.

e. User Response Evaluation: The employee attempts to answer the question. The LLM evaluates the response, providing feedback with grades and the correct answer.

f. User Inquiry and RAG Support: Employees can request explanations about any question from the pdf. The LLM utilizes RAG to retrieve relevant information and provide comprehensive explanations or summaries.

g. User Feedback: Employees can provide feedback on the fly.

h. Data Storage and Analysis: All user interactions, responses, grades, context, and feedback are stored in the database (db) for SRS logic (deciding the next question to ask) and for continuous system improvement. The context in db also serves as evidence of LLM being faithful.

I. Fine Tuning & Prompt Engineering

Since fine-tuning the Gemini API isn't currently possible, we've chosen to utilize prompt engineering for this initial prototype. Additionally, we've instructed the model to respond with "I don't know" when it lacks confidence in its response. This ensures that only questions with answers verifiable from the provided context are stored or shown to the user, enhancing the model's reliability and reducing the risk of hallucination.

QAG Prompt

"""You are tasked with generating questions from the following context. The answers to the questions should be found in the context.

Output the questions each on a new line.

Context: '''{context}''' """

User Response Evaluation Prompt

"""Given the following context, answer the question below.

If the context does not contain the answer to the question, output "I don't know."

Context: '''{context}'''

Question: '''{question}''''"""

IV. DISCUSSION

We successfully developed two prototypes in short time with minimal resources, thanks to the Gemini API and Microsoft Phi2 model which are freely available for public use (as of writing this article) and is straight forward. While many are available, we just chose one API type and one local CPU type.
Microsoft Phi-2 (2.7B local model) - This Phi2 2.7B model works on an even CPU. The out-of-the-box results were not that great as they gave liberal grades, even if the answer provided by the employee was not exact. We need to study further whether fine-tuning could fix this issue. To prioritize simplicity and avoid fine-tuning within constrained compute resources, we recommend opting for API-based services over local implementations. The API is faster in response as well.
As the most common training material are in the form of pdfs, we start with them. We fed seven pages from 'Python DA Fabio.pdf' to Gemini API.
We will stick to Gemini API.

A. Question-Answer pair generation (QAG)

Sample Question Answer Generation from pdf.
Question 1: What are the common formats for storing and collecting data?
Answer: XML, JSON, or simply XLS or CSV files

Question 2: What is the purpose of machine learning?
Answer: To recognize patterns, clusters, or trends and then extract useful information for data analysis in an automated way.

Question 3: What are some of the most commonly used statistical techniques in data analysis?
Answer: Bayesian methods, Regression, Clustering

Evaluation of the Model for QAG.

Performed human evaluation on the model, based on 100s of records. Gemini API works well out of the box. The following results seem satisfactory for our QAG and user response, as we are getting 90% + accuracy.

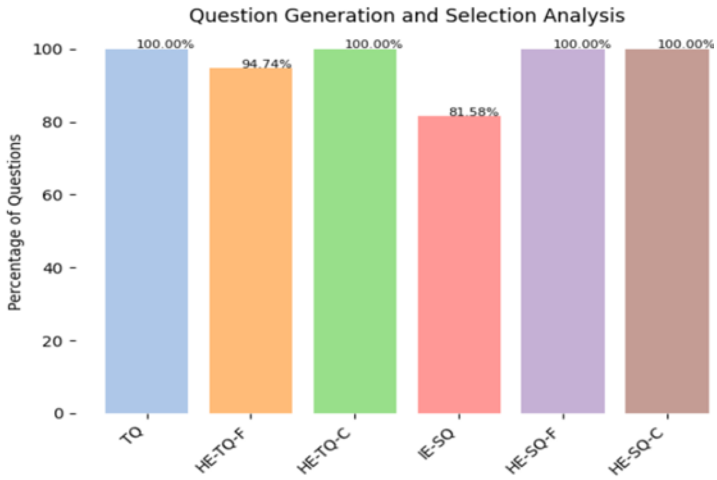

Figure 6 Gemini-based Model performance on QAG.

TQ: Total Questions generated by the BOT.

HE-TQ-F: Human Evaluation of the Total Questions for Faithfulness.

HE-TQ-C: Human Evaluation of the Total Questions for Correctness.

IE-SQ: Indirect Evaluation by the system to find good questions & remove others.

HE-SQ-F: Human Evaluation of the Selected Questions for Faithfulness. HE-SQ-C: Human Evaluation of the Selected Questions for Correctness.

B. User-response grading and feedback by the Bot

Evaluation for grading & feedback & RAG

Performed human evaluation on the model used for user response grading & feedback and found the results were 90% +, quite enough for our purpose.

Bot-1 asks a question from pdf:
What is required for parsing and extracting data from text files, web pages, and charts?
Employee's answer: Specific technical expertise for web scraping

Bot-1 response:
Grade: 8 out of 10

Feedback:
The answer is correct and concise. It accurately identifies the specific technical expertise required for parsing and extracting data from text files, web pages, and charts. However, the answer could be improved by providing a brief explanation of what web scraping is and how it is used in data analysis.

Correct answer:
Specific technical expertise for web scraping.
Context:
Specific technical expertise for the parsing and the eventual extraction of these data (called web scraping).
Source:
P1 Python DA Fabio.pdf, Page # 4

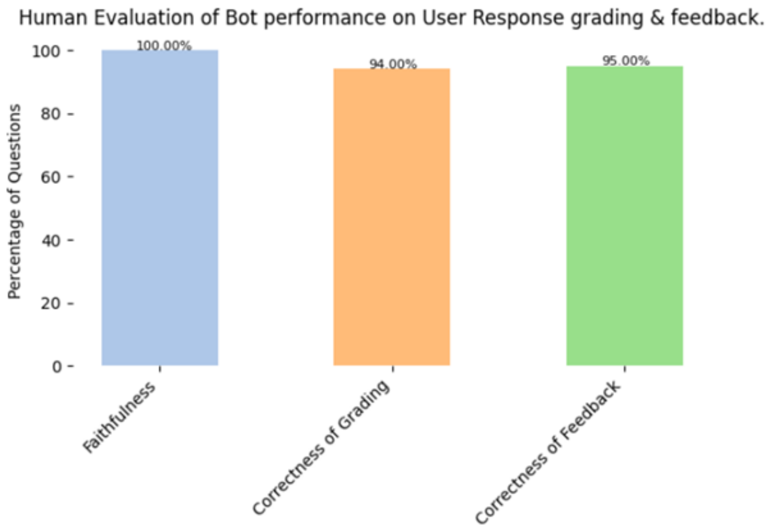

Figure 7 BOT's User response grading and feedback.

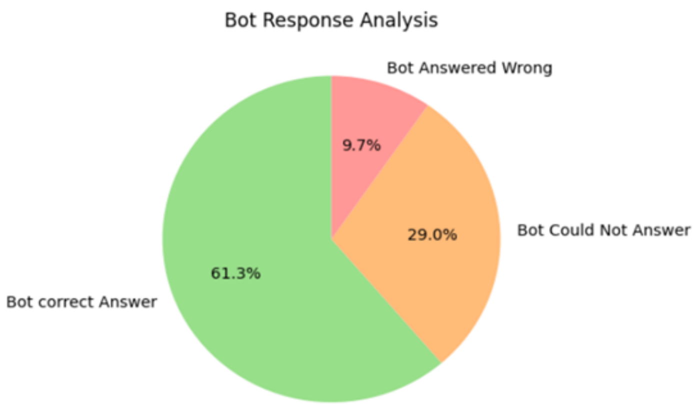

Figure 8 RAG performance.

Performed human evaluation on the model used for RAG and found the results were 60% + only; we still need to do some more research on the RAG part.

Module	% Correct
Questions filtered by the Self check	100
User answer grading & feedback	95
RAG	61.3
Overall accuracy	**85.43**

Table 1 Overall accuracy of the model

C. Challenges

LLM Bias: LLMs trained on biased data may generate biased questions or explanations, requiring careful data selection and mitigation strategies. This will be further studied and taken care of. However as we are confined to the target data, LLM bias might have less effect.

Overreliance on Technology: Overdependence on the system could hinder employees' development of independent learning skills.

Ethical Considerations: Data privacy and potential misuse of organizational data necessitate robust ethical frameworks and employee consent mechanisms.

Impact on Human workers: The OWCL is not replacing post-training follow-ups; it is just complementing them. In most organizations, as such activities are not happening at all, OWCL could be of immense use. The OWCL helps ensure that post-training follow-ups are done consistently and accurately.

While employees with busy schedules might see additional learning as a burden, offering incentives like prizes can help make it more engaging and worthwhile.

Technological limitations & Organizational resistance – Many organizations may have legacy systems, but they may have challenges in connecting to them. Also, the staff needs to be convinced of the advantages of the OWCL to avoid resistance.

Data Privacy: More research is required to implement a local model, such as office emails and meeting transcripts. The integration of these into the system might need more discussion due to data privacy issues.

V. FUTURE WORK

The production system will be implemented using programming languages and frameworks appropriate for the chosen architecture and functionalities: AWS or Azure & Hugging Face.

Advanced Preprocessing is to be further explored to suit a variety of input formats.

Continuous learning APIs will be explored for integrating with existing systems, such as daily work reports, meetings transcripts, email bots etc. for the system to upgrade itself on those topics. Also, the admin or users could add new updates to the existing courses.

Context-aware SRS Integrates a machine learning (ML) algorithm to analyze question and response context within the SRS, preventing repetitive questioning and promoting deeper understanding through similar questions within the same topic. One of our co-authors have already published a research paper on a different use case [17], which could be further explored and modified accordingly.

Provision for the user to provide feedback on each question's quality while in production.

Analyze user interaction data and performance metrics to assess the effectiveness of the system's learning reinforcement strategies and identify areas for improvement.

User privacy will be ensured by adhering to relevant data protection regulations and anonymizing collected data whenever possible. User consent will be obtained for data collection and usage.

The recommender system depends on the historical data of all the employees' revision and recall grades and prepares reports, which are commonly difficult topics, as well as personalized reports. The system would suggest further which trainings were successful and which need to be repeated.

Proctored exams could be implemented using the same system, which provides more authority to the user's grades.

Investigate the long-term impact of such systems on student learning habits and academic performance. Explore advanced user interface designs to enhance user engagement and interaction with the system.

VI. CONCLUSION

We successfully developed a prototype for a novel idea, Organization-Wide Continuous Learning (OWCL), which utilizes personalized AI chatbots powered by large language models (LLMs) to enhance post-training knowledge retention. This system achieved impressive accuracy (over 94%) in core functionalities like question generation and answer evaluation. Our results demonstrate the significant potential of LLMs to revolutionize post-training learning by improving employee knowledge retention and engagement. With an overall 85% accuracy, our research suggests that OWCL is a balanced and practical approach, leveraging cutting-edge technologies like APIs of very large language models (vLLMs) while remaining accessible, resource-efficient, and fast.

However, the RAG part needs some more research. We also experimented with a small model – Phi2, the out of box results were not very impressive.

The above amazing results were received upon implementing the Gemini Pro API, which is a very large language model that was recently released by Google.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my peer mates at Bits Pilani (Work Integrated Learning Program) Praveena Parameswaran & Veena Vijayan for their support throughout this research project. Their insightful brainstorming sessions during the earlier phases of the research were invaluable and helped shape the direction of this project. Chandan Maity’s & Nalin Chandra Giri’s insightful feedback on the manuscript was greatly appreciated. We are grateful to the WILP Division (Data Science) at BITS Pilani for providing the support necessary to complete this research. We would also like to thank our professors for their invaluable guidance throughout this project, and the office staff for their support.

REFERENCES

[1] H. Ebbinghaus, "A Contribution to Experimental Psychology," New York by Teachers College, Columbia University, p. 140, 1885.

[2] K. Rivera, S. D. Sala, A. D. Baddeley and R. Logie, "Rate of forgetting is independent from initial degree of learning across different age groups," ResearchGate, p. 11, 2022.

[3] N. J. Yael Wollstein, "Spaced Effect Learning and Blunting the Forgetfulness Curve," ResearchGate, p. 10, 2023.

[4] H. J. Martin, "Improving Training Impact Through Effective Follow-Up: Techniques and Their Application," Journal of Management Development, p. 16, 2010.

[5] J. K. Ford, T. T. Baldwin and J. Prasad, "Transfer of Training: The Known and the Unknown," Annual Reviews, p. 27, 2018.

[6] M. I. Hanif, H. Sagheer and S. Mehak, "Impact of Knowledge Sharing and Knowledge Retention onEmployees Development through the Mediation of SustainableCompetitive Advantage and Moderation of TransformationalLeadership," ResearchGate, p. 14, 2020.

[7] Google, "Gemini: A Family of Highly Capable Multimodal Models," arxiv, p. 62, 2023.

[8] Bubeck, M. Javaheripi and Sebastien, "Microsoft Research Blog," December 2023. [Online]. Available: https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

[9] S. Sreelakshmi, S. B. Abhinaya, A. Nair and J. Nirmala, "A Question Answering and Quiz Generation Chatbot for Education," ResearchGate, p. 7, 2019.

[10] X. Yuan, T. Wang, Y. H. Wang, E. Fine, R. Abdelghani, P. Lucas, H. Sauzeon and P. Y. Oudeyer, "Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation," arXiv, p. 12, 2022.

[11] S. Huo, N. Arabzadeh and C. Clarke, "Retrieving Supporting Evidence for Generative Question Answering," ACM Digital Library, p. 10, 2023.

[12] Olney, "Generating multiple choice questions from a textbook: LLMs match human performance on most metrics," p. 18, 2023.

[13] J. K. Matelsky, F. Parodi, T. Liu, R. D. Lange and K. P. Kording, "A large language model-assisted education tool to provide feedback on open-ended responses," arXiv, p. 7, 2023.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. T. Yih, T. Rocktaschel, S. Riedel and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Task (RAG)," arXiv, p. 19, 2021.

[15] D. J. Wittman, "StanislousState," [Online]. Available:

[16] https://www.csustan.edu/sites/default/files/groups/Writing%20Program/forgetting_curve.pdf.

[17] A. Chun and H. H. Ja, "The effect of flipped learning on academic performance as an innovative method for overcoming ebbinghaus' forgetting curve," ResearchGate, p. 5, 2018.

[18] N. Bhore and S. Page, “Innovative Approach for Application of Machine Learning Techniques in Mobile Payment Fraud Analytics,” vol. 27, no. 27, pp. 33–48, 2020.