Reimagining Tacit Knowledge Extraction and Summarization by Inferencing Large Language Model in a High-Performance Computing Platform

Rahul Hardikar
Product Specialist
TCS R&I
Vadodara, India
hardikar.rahul@tcs.com

Saurabh Barve
Product Owner
TCS R&I
Pune, India
s.barve@tcs.com

Shampa Sarkar
Scientist
TCS R&I
Mumbai, India
shampa.sarkar@tcs.com

Revati Kulkarni
Technology Head
TCS R&I
Pune, India
revati1.kulkarni@tcs.com

Abstract*— Tacit knowledge that has been gained over several years is critical for product R&D of an organization. This data is unstructured and multi-modal in nature, consisting of images, documents, and notes that are written in domain specific language. In the automotive industry, a design release engineer (DRE) is responsible for generating design documents from such diverse data/tacit knowledge that is spread over departments and years. Any inaccuracies in the design document can lead to product recalls, resulting in revenue losses of millions of dollars and a tarnished brand image. Our solution accelerates this pipeline from data ingestion from multiple sources to templatized and accurate, contextualized document generation. It optimizes the business process by reducing the time from weeks to mins.

Keywords—Heterogeneous Orchestration, High Performance Computing, Generative AI, Large Language Model (LLM)

I. INTRODUCTION

Tacit knowledge or explicit knowledge is unstructured in nature. It can be in the form of text, image or mathematical formulae, as data or metadata, distributed or embedded in several hundreds or thousands of documents. Tacit knowledge can also be obtained in the form of domain specific language (DSL), which is different from general-purpose language (GPL).

Currently, tacit knowledge extraction or summarization is a manual process, specifically in the context of statements of requirement (SoR) generation by Design Release Engineer (DRE) in the automotive industry. This process is tedious, human error-prone and complex. An approach towards automating the process is by the natural language generation (NLG) using large language models (LLM). However, the challenges existing are manifold, for example hallucination in the response generation by the LLM, obtaining low latency response generation towards knowledge extraction and summarization, etc. Conventional methods also suffer from lack in contextuality to subject matter specifically in a mixed scenario of domain specific language and general-purpose language, and in accuracy of output/response with respect to human alignment. There exist several algorithms and methodologies towards response generation with human feedback. However, low latency, contextual and accurate tacit knowledge extraction and summarization are challenging tasks.

II. MATERIALS AND METHODS

The extraction and summarization of the information content is developed at scale (achieving latency reduction with accuracy) using an AI/ML assisted high performance computing (HPC) platform (TCS A3 – TCS Accelerated AI and Analytics) [1]. The heterogenous HPC architecture of the platform leverages large language model (LLM) through a retrieval augmented generation (RAG) pipeline, wherein each component of the heterogeneous RAG workflow is orchestrated optimally (for e.g., with respect to selection of the vector DB, selection of the LLM towards response accuracy, selection of the guardrail with topic toxicity marker/remover, selection of hallucination module towards response generation and the inference module towards reduced latency and high throughput, etc.).

For the RAG workflow, all the relevant contexts to a given complex query are retrieved using a novel payload indexing schema, while querying the vector database comprising multimodal knowledge documents. All the retrieved contexts are chained by context-chaining to concatenate with the given prompt towards summarization of the extracted tacit knowledge response in a few-shot in-context learning mode. A query from DRE, such as make, model, year, part, can provide multiple contexts. Our novel workflow does dynamic payload indexing to ensure all the relevant documents matching the query are provided as contexts to the LLM. Further we have introduced guardrails and a hallucination module to ensure precise and accurate output. Time to generate an output scale linearly with the number of contexts provided to the LLM. This process is optimally parallelized using a scalable distributed heterogeneous computing engine resulting in latency reduction of 10x for 1 GPU and 30x for 8GPUs. The dynamic batch scheduling engine of TCS A3 makes the workflow scalable. It is a low-latency, contextual

tacit knowledge summarization solution that can save an annual effort of 100-thousand-person hours.

III. RESULTS AND DISCUSSION

Figure 1 provides the solution workflow that can assist towards the tacit knowledge extraction and summarization by the DRE. We have built a tacit knowledge vector database (using Qdrant vector database) comprising a variety of data and metadata of text, image, mathematical formulae, where the text and image metadata (text) are tokenized, and vector embedded (dimension d = 384) using *all_manylm_lc_v2 tokenizer*. The vector database is also hosted on high-speed storage devices (parallel file system), where the capacity of the storage is expandable based on use cases. Distributed data parallelism (DDP) technique is further tuned for the parallel vector embeddings, which helps in lowering the process latency significantly.

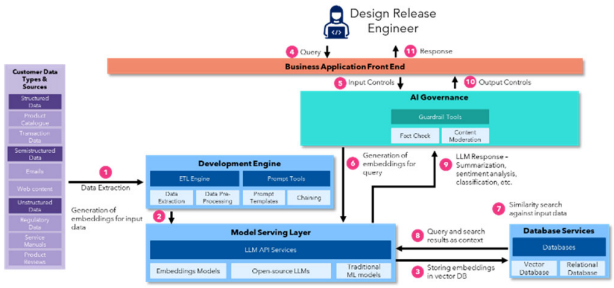

Fig. 1. The Solution Workflow of Tacit Knowledge Extraction and Summarization

We present here a low latency inference engine integrated with the LLM solver, which is configured to generate responses which are reduced by several orders in latency with respect to industry standard inference engine [2] (as shown in Figure 2). For standard inference, total inference time increases on the number of prompts being run. For low latency inferencing, the inferencing is performed by dynamic batch job scheduling of the plurality of prompts and further incorporating distributed data parallelism techniques. As shown in Fig. 2, a 10-fold reduction in the inferencing with low latency inference engine is achieved on a single GPU and a 30-fold reduction on a multi-core 8 GPU is achieved.

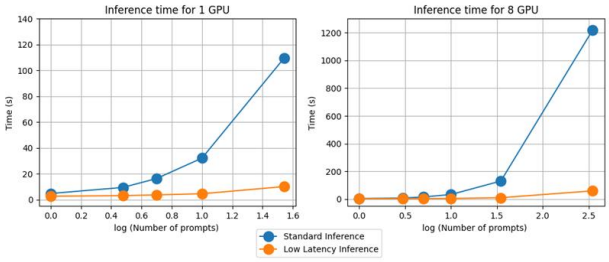

Fig. 2. Reduction in Inference Latency Towards Tacit Knowledge Summarization

IV. CONCLUSION

The solution provides low-latency, contextual tacit knowledge extraction and summarization solution for multimodal structured or unstructured data. The document generated is accurate and hence eliminates human error that can result in negative business impact. The work incorporates distributed training and inferencing using the high-performance computing platform of TCS A3 which reduces the latency significantly.

V. ACKNOWLEDGEMENT

The authors acknowledge Abhilash Adavi for solution development and Shashi Bhushan for technology review.

VI. REFERENCE

[1] *https://www.tcs.com/who-we-are/newsroom/press-release/tcs-establish-advanced-computing-center-high-performance-computing-solutions*

[2] *GitHub - bentoml/OpenLLM: Run any open-source LLMs, such as Llama 3.1, Gemma, as OpenAI compatible API endpoint in the cloud*