A community initiative, anchored by

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

IIIT Hyderabad

TechForward

DISPATCH

AUGUST EDITION

FOUNDATIONAL AND LARGE LANGUAGE MODELS

Foundational models serve as versatile tools for AI and industry, providing a base for a wide range of applications by understanding and generating human language. Large language models, a sub-set of foundational models, leverage their extensive scale and training to deliver sophisticated capabilities like content creation, customer support, and data analysis. The TechForward August edition explores how these models are driving innovation by enhancing automation, improving decision-making, and offering personalized user experiences. Their impact is evident across sectors, from tech and finance to healthcare and entertainment, reshaping how businesses interact with and leverage data.

IIITH's TechForward research seminar series is an academia-industry confluence around emerging technologies. The deep insights, directional talks and industry outlooks from accomplished thought leaders at the seminar are compiled monthly in the Tech Dispatch as a ready reckoner for technology directions.

TECHFORWARD AUGUST EDITION SNAPSHOTS

CONTENTS OF THIS EDITION

Indic LLMs and LLMs in Education

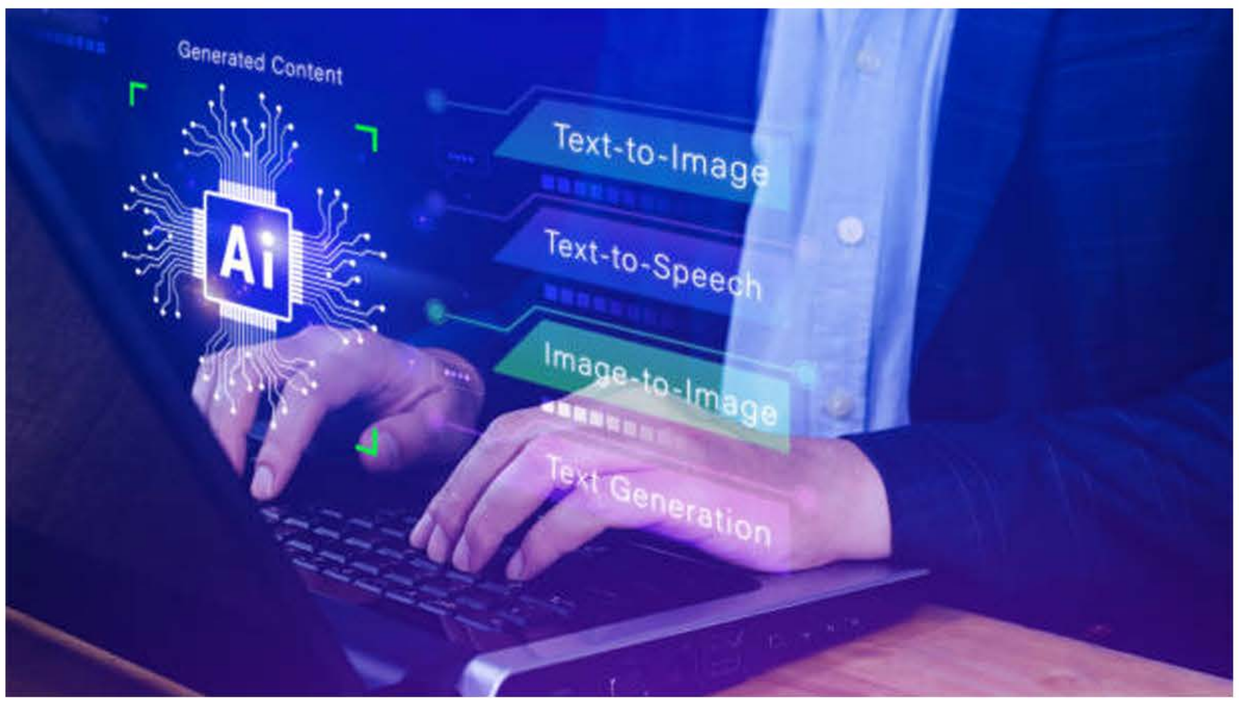
***Prof. Vasudeva Varma** explores two use cases of Generative AI and LLMs - in the Wikipedia development in Indian languages and in education and learning. This article is based on key parts of his talk from the TechForward Research Seminar Series on Foundational and Large Language Models.*

GenAI and LLMs are the latest buzzwords in the technology world. They are used interchangeably but in fact are two distinct technologies. GenAI can be defined as artificial intelligence capable of producing original content, like text, images, music and so on. LLMs or large language models on the other hand, are a specialised class of AI models that use natural language processing to understand and generate human-like text. Both GenAI and the LLMs have disrupted the world immensely. The world of LLMs is changing dramatically with the frantic pace of innovation. It has also caused a lot of confusion among researchers from academia on what problems to work on because LLMs seem to be solving everything. Indian LLMs or Indic LLMs hold a lot of contextual and cultural relevance and it will be interesting to see how the efforts of BharatGPT, Sarvam, Krutrim and so on will unfold.

Most of the progress in AI so far has been on the part of the iceberg that is above the water. There is yet a lot of progress to be made on the part that is below the water. The biggest opportunity for India is to look at the sectors where the Indian economy is already strong and figure out the applications of AI to those sectors that enable it to maintain its advantages. Here's a look at the two main use cases from India where our efforts have yielded a bigger impact: Wikipedia development in Indian languages and adopting LLMs in education and learning.

Wikipedia

Just like ancient civilizations were dependent upon and developed along the banks of great rivers, our modern society too depends on a 'river' of knowledge flowing through our language and culture. The 'river' in question today is Wikipedia. The existence of Wikipedia is crucial for everything we do. But it's not just Wikipedia in isolation; there's an entire Wiki ecosystem comprising Wiki Data, Wiki Source, Wiki Commons and other projects within Wikipedia that leverage each other and that is the combined power and synergy of Wikipedia.



As a Cultural Tool

In its potential as a 'knowledge river', Wikipedia's presence in Indian languages however is currently inadequate. While the English language has more than 7 million articles on Wikipedia, the best of the Indian languages in terms of representation on Wikipedia, Hindi has around only 150,000 articles or so. At IIIT-H, we have set out to correct this anomaly and enhance Indian language content in Wikipedia with the IndicWiki project. Creating encyclopaedic content is a very cognitively-demanding task involving a lot of research, collection of relevant material and references before the article can be penned. In addition to this, the material should conform to the 'five pillars' which includes writing from a neutral point of view, in a manner that anyone can use, edit and distribute the content and with the underlying premise of editors treating each other with respect and civility.

Telugu Wikipedia Content

We came up with a two-fold solution for this: One is template-based development of Wikipedia articles and the second is automated generation of these articles. The first solution starts with creating a structured manual. We invited people who can read and write in their mother tongues like Telugu, Oriya, or Hindi really well. They were also expected to have the knowledge of Python. So equipped with Telugu and Python, and with the help of the manual, they were able to create 1000s of articles in a particular domain. Multiple data sources brought together large amounts of data and within 6 weeks, we were able to create around a million articles in Telugu and about 200,000 articles in Hindi in each of those domains. While this approach was a great success in itself, we however didn't want to stop at that. We explored the generation of encyclopaedic articles with the help of newer models and approaches using Generative AI. This is our second solution.

We looked at three sub problems to automatically create a credible Wikipedia article in Indian languages. One, we created a cross-lingual dataset of aligned facts with text. It means that for each fact that exists in Wiki Data, we have the mapping of the text in any of the wikipedia articles in any of the Indian languages. Such a data set is being used to create a model which in turn is being used to generate text in any language from any new fact that is published. So, we've done a cross-lingual fact-to-text-generation of encyclopaedic sentences. It gives us shorter pieces of text that are worthy of being in an encyclopaedic article. The other problem we tried to solve was, given a set of documents on a specific topic, we explored the creation of an outline of an article that can go into Wikipedia. So essentially from the two sub modules, there is a skeleton of an article that is being created. The third problem for which we came up with a solution relates to the references that are present at the end of the articles. We tried to augment the outlines of the articles from the references. The foundational large language models could not help us much here, so we had to use an innovative and new kind of approach to generate encyclopaedic articles in Indian languages using resources coming from the Wikipedia ecosystem.

GenAI and Learning

There are experiments in the learning space too by using GenAI. This was demonstrated in a mobile app-based course developed by ISB for the Rajiv Gandhi University of Knowledge Technologies (RGUKT) for bettering the English-speaking abilities of the RGUKT students who hail from a rural background. The way it works is that a text is displayed on the screen prompting the learner to read it out aloud. As he or she reads it out, the voice is recorded and analysed. The app then gives feedback in terms of mispronunciation, omissions from text, additions to original text and so on. It was found that by playing with the app repeatedly, the students were able to gain a lot more confidence in speaking English. Similarly there is an AI model that powers a negotiation course taught by ISB which again has become very successful. This model acts as an agent with whom the learner needs to negotiate to close a given deal. They are looking at enhancing such models and tools in other areas of management education. However, there are issues of scalability and adaptability that are being worked upon.

Middleware For Education

When there are LLMs on one side and learners on the other, one way to bridge the gap between the two is to build middleware for it. The idea is also to make the middleware useful for teachers for instructional design. The other thought is to make it a conversational tool that improves meta-cognitive skills, and enhances accountable talk - free-flowing discussions in the classroom. But along with the middleware, there come responsibility layers too. A great example lies in Khanmigo, a bot created by Khan Academy. It patiently guides the students and helps them arrive at the right knowledge levels not by giving them answers but by letting them use their own reasoning abilities in coming up with the required answers. Another example is from OpenAI where they're currently looking for model teachers who will train the models with the right kind of behaviours. Similarly, Google's LearnLM is a family of models fine-tuned for learning and teaching experiences. It helps in reducing the cognitive load while watching educational content, ignites curiosity, adapts to the learner and deepens metacognition. There are other new tools such as the Illuminate, which use the LearnLM model. Illuminate breaks down research papers into short audio conversations providing an overview of key insights from the papers. One can also chat with the model and ask follow-up questions. Another tool is the "Learn About" experience. Here again, one can quiz the model and it helps guide you through any topic at your own pace. All of these are excellent examples of how foundational models can be fine-tuned to meet the demands of education and learning. However, technological innovations notwithstanding, tech will always be incident to the learning process.

No technology can ever replace the magic of a teacher, but when applied in deliberate and thoughtful ways can help augment the teacher's capacity, giving them time to invest back in themselves.

Conclusions

The key point I want to make is that the last mile journey in the GenAI application is the hardest. While technology may enhance productivity, domain knowledge is the cornerstone to success. For example, if classroom insights from teachers and students are not captured, then we will not be able to create truly impactful learning experiences.. And finally for these foundational models, one needs to create better responsibility layers before applying or adopting in a domain.

PROF. VASUDEVA VARMA
IIIT HYDERABAD

is the Head of the Language Technologies Research Centre at IIITH. His research interests are in the broad areas of information retrieval, extraction and access and more specifically social media analysis, summarization, semantic search, text generation, and cloud computing.

Language Tech Through The Ages

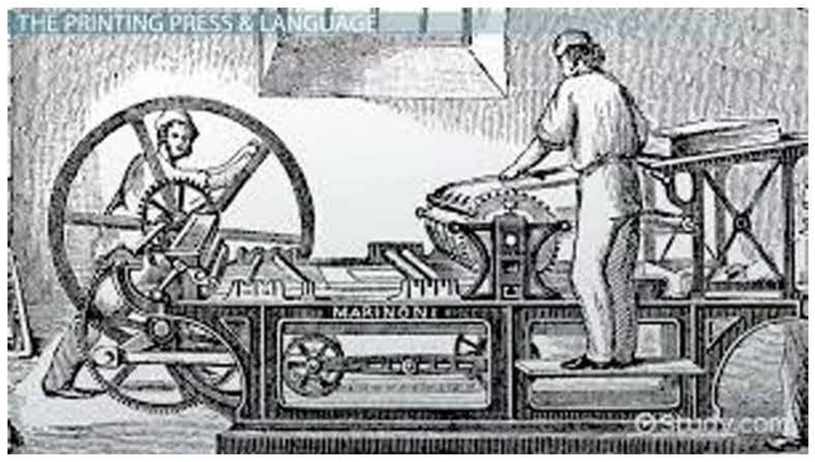
***Prof. Dipti Misra Sharma** traces the evolution of language and language technologies before summarising the research developments in the space that are taking place at IIITH.*

Perhaps the single-most distinguishing feature of humankind is its ability to speak. While other species communicate too, what sets humans apart is a far more advanced and complex system of linguistic signs that enables formulating and expressing abstract thoughts and ideas beyond mere speech. 'Language' has played a very critical role in the way humans have progressed in the realm of knowledge sharing and preservation leading to human advancement. Language is not merely a tool for communication. It is a vehicle for culture, identity, and social cohesion. Language evolved as a spoken system. But language only in the spoken form had limitations. Knowledge dispersal and long distance communication were difficult in this mode.

One of the major landmarks that facilitated knowledge/information dispersal was the invention of writing systems. Writing allowed humans to preserve their thoughts and ideas facilitating communication across time and space. From painstaking carvings on stone, 'writing' moved to imprinting text on clay tablets and using reeds dipped in ink before modern-day pens and pencils were invented. But this still meant that the written word was available to only a select few. It wasn't until the invention of the printing press that widespread dissemination of the written word took place.

Early Tech Advancements

The introduction of the printing press in the 15th century revolutionised the dissemination of written content. It enabled the mass production of texts, making books more accessible and affordable. What truly brought about a major technological shift is the advancements in natural language processing which moved from mere representation of the spoken language towards an effort to perform some language functions that involved understanding and generation of human language in the manner that humans do.



The task was extremely challenging. Modelling language in a way that can emulate human use of language for communication is not easy. Communication using language by humans requires several types of knowledge. It includes not just complex linguistic knowledge but also knowledge of prosody, facial expressions, gestures, shared experiences, knowledge of the world around and so on. It has taken decades (Second world war to 2024) for language technologies to evolve to the level we are at today since incorporating the kind of knowledge that humans have for language use is an extremely challenging task. Initial efforts in NLP (largely for machine translation) were very limited, relying on basic dictionaries and some grammar rules. These were fragile and not really very usable in a real-world scenario.

Evolution of NLP

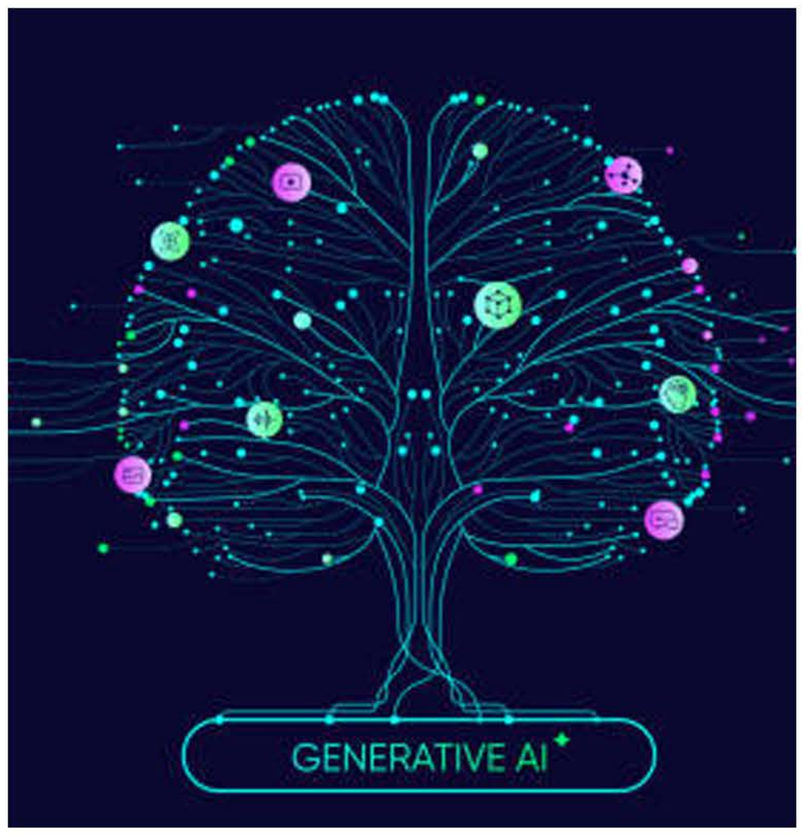
At the beginning of the Cold War, the IBM 701 computer automatically translated Russian sentences into English for the first time. In the 60s, the creation of ELIZA - one of the first chatbots to simulate conversation - marked a notable achievement in the field of NLP. It was in the 70s that the first ideas around rule-based machine translation emerged.The 1980s marked a shift towards more complex language processing systems, with the introduction of conceptual ontologies that structured real-world information into computer-understandable data. This shift allowed for the development of models that could learn from data rather than relying solely on predefined rules. This lead to the emergence of statistical methods as a solution to overcome the limitations of rule-based systems for training models. The introduction of word distribution semantics further advanced NLP by emphasising the relationships between words based on their context. Long Short-Term Memory (LSTM) networks emerged in the late 1990s, revolutionising the processing of sequential data in NLP. This capability was crucial for tasks such as speech recognition, language modelling and text classification. All these advances were small steps forward to model natural language for doing complex tasks.

The introduction of the Transformer architecture in 2017 marked a significant leap in the field of NLP. Transformer-based multilingual approaches have shown processes for multiple language tasks including Indian languages like Hindi, Bengali, and Gujarati, leading to substantial advancements in language technology. Speech-to-speech machine translation is another transformative architecture built by combining multiple different language technologies that has the potential to effectively address multilingual needs. Today, the talk is all about large language models (LLMs). The rise of LLMs is based on advances in the use of large text data and computational resources. However, development is biased towards high-resource languages like English, underrepresenting many languages, including several Indian languages. The lack of representation can lead to models that aren't culturally sensitive or which perpetuate biases, a concern in diverse countries like India. Another concern that the advent of this technology and its heavy reliance on large data and huge compute facility poses is introducing disparities leaving languages behind which do not have a reasonable digital presence. If a language is left behind, then it is not just the language but also the societies, their culture and knowledge of these cultures that are left behind. Hence, going forward, the technology will move towards addressing these concerns.

In India’s multilingual context, these technologies can make a vast difference to people’s life in critical domains of health care, judiciary and education, apart from meeting people’s day today communication needs. Collaborative efforts like the National Language Translation Mission aim to create a more inclusive digital landscape by developing resources and technologies for Indian languages. Academia, industry, and government collaborations can pool resources to tackle these challenges. Encouraging open-source datasets can enhance training data availability for LRLs, enabling researchers to build better models without prohibitive data acquisition costs.

Language Tech at IIITH

In this evolving landscape of language, IIIT Hyderabad has been playing its role by developing language-related resources and technology and by being part of major national projects in this area. The effort spanned from the era of rule-based systems to statistical methods and recently to LLMs tailored for multiple Indian languages. The work includes a wide range of applications starting from core language understanding and generation tools such as shallow parsers, state-of-the-art machine translation models, question-answering, summarization, dialogue processing, and speech-to-speech translation systems. In addition, this focused research has resulted in the development of various types of multilingual content, including the translation of educational video lectures, healthcare content, and legal materials along with end applications such as educational chatbots, allowing students to easily get answers to their educational questions in their native language.



The Future Of NLP

Research in language multimodality through the incorporation of various modalities like text, images, audio, and video-based factual and stored data will enhance comprehension, generation, and execution of language depending on the wider context. Future research in this will likely focus on developing more sophisticated algorithms that can seamlessly fuse these modalities, enabling machines to derive richer insights and more nuanced interpretations of content and aid us more in our day-to-day activities. Combining visual elements with textual descriptions can significantly enhance applications in fields such as human-computer interaction. For example, a multimodal chatbot that uses natural language processing (NLP) and video processing to help citizens access social welfare resources and services. Citizens can interact with this chatbot through both audio and visual means, asking questions regarding social services, including agricultural assistance and social schemes, in domains such as governance or health. Furthermore, users can upload images or videos as part of their inquiries or provide information-based proofs, which the chatbot can interpret to offer suitable multimodal responses. This approach not only improves user engagement, but also ensures that community members receive timely and pertinent information, leading to a more informed and cohesive community.

Conclusions

Language technology has come far in the last few decades. However, there is still a long way to go. The technology has reached a stage where it can be used to support some of the human tasks but it is still far from being anywhere close to the creative and social aspects of language use. Language is a critical component of human intelligence, thus an important field in AI research where we have to keep delving deeper to understand more complex aspects of the human mind and its linguistic abilities.

PROF. DIPTI MISRA SHARMA
IIIT HYDERABAD

is a Professor Emeritus at IIITH. Her areas of research interest include Machine Translation and Linguistics.

Large Language Models and Retrieval Augmented Generation: A Landscape for Precise Information Extraction

***Prof. Manish Shrivastava** explains how Retrieval Augmented Generation (RAG) optimises output of large language models and improves the quality of their responses.*

Large Language Models (LLMs) have emerged as powerful tools for processing and generating human-like text. These models, such as GPT-3, BERT, and their successors, have revolutionised natural language processing tasks including text generation, translation, and question-answering. Their ability to understand context and generate coherent responses has made them invaluable in numerous applications across industries.

However, when it comes to precise information extraction, especially at an enterprise scale, LLMs face certain limitations. One of the most significant issues is their tendency to "hallucinate" or generate plausible sounding but factually incorrect information. This occurs because LLMs are trained to predict the most likely next word in a sequence, rather than to retrieve and present factual information. While this approach works well for general language tasks, it can lead to inaccuracies when precise, up-to-date information is required. This is where Retrieval Augmented Generation (RAG) comes into play, offering a solution that combines the strengths of LLMs with the accuracy of retrieval-based systems.

Retrieval Augmented Generation

Retrieval Augmented Generation addresses the limitations of traditional LLMs by combining them with information retrieval systems. RAG models work by first retrieving relevant information from a knowledge base and then using this information to guide the language model's generation process. This approach significantly reduces the likelihood of hallucinations and ensures that the generated content is grounded in factual, retrievable information. It also allows RAG systems to leverage the vast knowledge and language understanding capabilities of LLMs while maintaining a high degree of accuracy and reliability in the information presented.

Impact on Enterprise-Scale Information Extraction

The combination of Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) is poised to revolutionise information extraction processes at the enterprise level. Traditional methods often struggled with context-dependent nuances and semantic complexities inherent in human language. When LLMs are coupled with RAG's ability to ground responses in verified information from curated knowledge bases, they result in highly trustworthy results. This enhanced accuracy translates directly to more reliable decision-making processes.

Customizability is yet another dimension where LLM-RAG systems shine in the enterprise context. Every organisation has its unique lexicon, domain-specific knowledge, and particular areas of focus. By curating a knowledge base that reflects an organisation's proprietary information, enterprises can ensure that the information extraction process is highly relevant to their specific context.

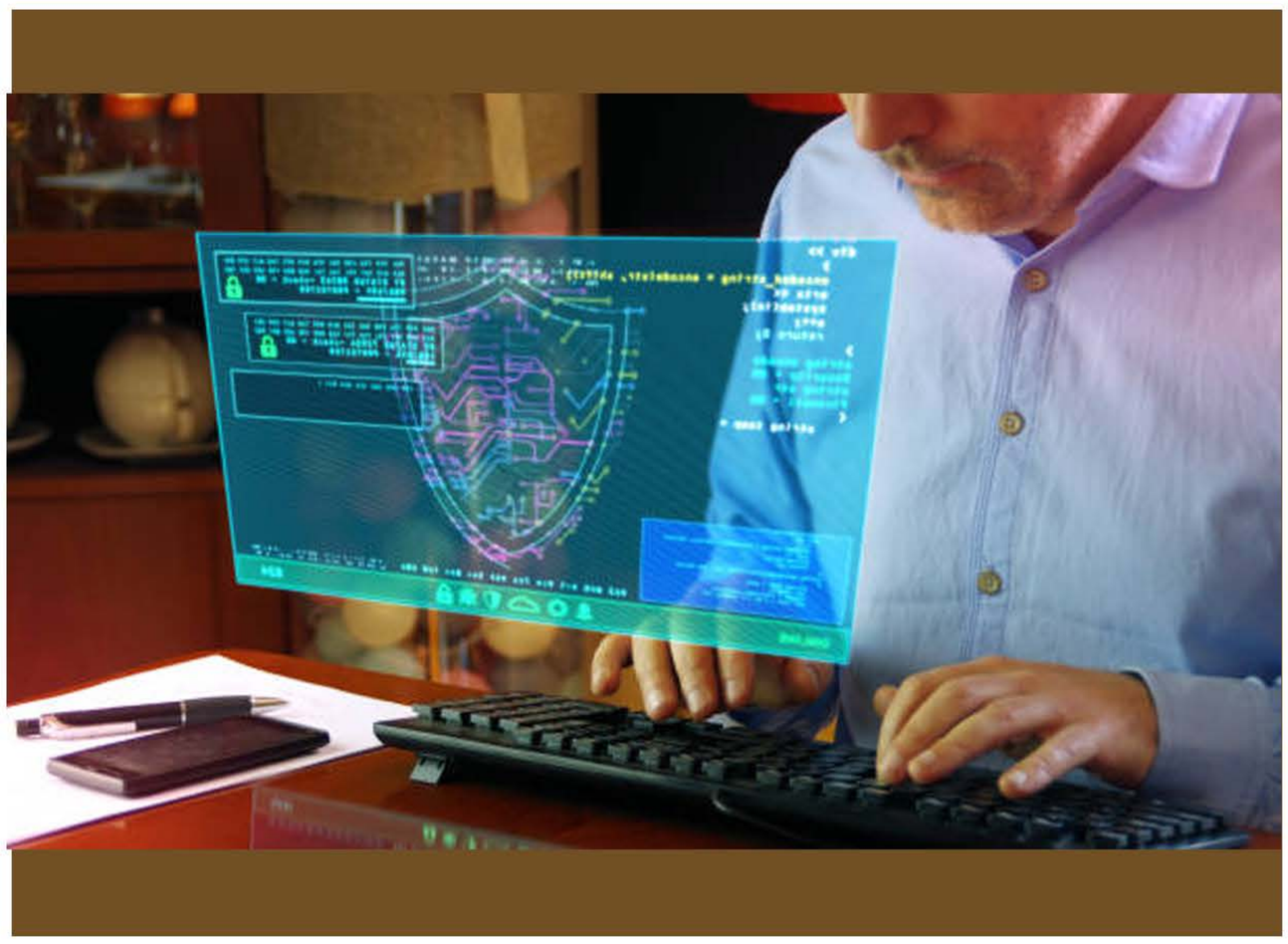
Transparency and explainability, often cited as concerns with AI systems, are significantly enhanced in LLM-RAG implementations. Unlike pure LLM systems, which can sometimes be opaque in their decision-making processes, RAG systems can provide clear references to the sources used in generating responses. This traceability is invaluable in enterprise settings for accountability and auditability.

Moreover, the impact of LLM-RAG systems extends beyond mere information extraction – it has the potential to transform how knowledge is disseminated and utilised within an organisation. By providing easy access to accurate, context-aware information, these systems can democratise knowledge across the enterprise. Employees at all levels can leverage these tools to quickly find relevant information, answer complex queries, and gain insights that were previously siloed or difficult to access.

RAG Offerings in the Market

The landscape of Retrieval Augmented Generation (RAG) solutions is rapidly evolving. The market is characterised by a mix of established tech giants, specialised AI companies, and open-source projects, each bringing their unique strengths to the table. Elastic, a company primarily known for its Elasticsearch search engine, has recognized the potential of RAG and integrated these capabilities into its offerings. Their approach cleverly leverages the powerful search and analytics capabilities of Elasticsearch, combining them with machine learning models to create a robust RAG solution. This integration allows for efficient retrieval of relevant information from large datasets, which can then be used to augment language model responses. Elastic's solution is particularly well-suited for enterprises that already utilise Elasticsearch for their data storage and retrieval needs, providing a seamless path to implementing RAG within their existing infrastructure.

Amazon's Kendra, while not explicitly marketed as a RAG solution, is an intelligent search service that can serve as a crucial component in a RAG implementation. Kendra uses machine learning to enhance search results and can be seamlessly integrated with other AWS services to create comprehensive RAG systems. This makes it a particularly attractive option for enterprises looking to improve information retrieval from their internal documents, websites, and databases while leveraging their existing AWS infrastructure.



Similarly, Microsoft's Azure Cognitive Search can be utilised as the retrieval component in a RAG system. It offers AI-powered search capabilities that can be combined with Azure's machine learning services to create end-to-end RAG solutions. Microsoft has been actively working on integrating RAG-like capabilities into its broader suite of AI services, enhancing their accuracy and relevance across various applications.

As the field of RAG continues to evolve, we're seeing an increasing focus on the underlying data structures that power these systems. Two key approaches have emerged as dominant in the market: vector databases and knowledge graphs. Each of these technologies offers distinct advantages and limitations, shaping the way RAG systems are implemented and influencing their performance in different use cases.

Vector Databases in RAG Applications

Vector databases have become a popular choice for many teams implementing RAG systems, primarily due to their ease of setup and the speed of retrieval they offer. These specialised databases are designed to store, index, and query vector embeddings - numerical representations of unstructured data such as text, images, and audio. The process of using a vector database in a RAG system typically involves several steps. First, the raw data is ingested and pre-processed, which includes cleaning the data and segmenting it into manageable chunks. These chunks are then converted into vector embeddings using an embedding model, which captures the semantic meaning and context of the data. These embeddings are stored and indexed in the vector database. When a query is made, it is processed through the same embedding model, and the resulting query vector is matched against the stored embeddings to find similar vectors. The retrieved data is then combined with the original query and passed to a large language model to generate a contextual and holistic answer.

However, vector databases do have limitations. They lack the ability to represent hierarchical or relational structures found in traditional databases, which can be limiting for use cases that require organised data relationships. Additionally, search results are often limited to the top several matches to avoid overwhelming amounts of data, which can hinder effectiveness in cases where more comprehensive data retrieval is necessary.

Knowledge Graphs in RAG Applications

While vector databases have been the go-to solution for many RAG implementations, knowledge graphs are gaining traction, particularly in enterprise AI spaces where understanding complex data relationships is critical for better accuracy. Knowledge graphs offer a structured way to organise data, representing entities as nodes and relationships as edges within a graph.

Implementing a knowledge graph differs significantly from vector database approach. First, data is extracted from various sources and processed to identify entities, relationships, and metadata. This information is then used to build the knowledge graph, creating nodes for entities, edges for relationships, and tagging all elements with associated metadata. When a query is made, the system identifies relevant entities and relationships within the query and constructs a graph query to retrieve the pertinent information. The retrieved data is then used to augment the context provided to the language model, enabling it to generate a more comprehensive and accurate response.

Advantages of Knowledge Graphs

The use of knowledge graphs in RAG systems offers several advantages. They can significantly improve retrieval accuracy by accounting for relationships between data points, leading to more contextually relevant responses. Knowledge graphs also provide transparency and data lineage by adding metadata to nodes and edges, making it easier to trace the origin and evolution of data. This enhanced explainability can be crucial in building trust in AI systems, especially in sensitive domains like healthcare or finance. However, just like vector databases, knowledge graphs also come with their own set of challenges. The setup and maintenance of knowledge graphs can be complex and resource-intensive, requiring significant effort in data modelling and ontology design. They also tend to be slower at retrieving data compared to vector databases, as they need to traverse the graph to answer queries. Additionally, handling data updates in a knowledge graph can be challenging, potentially leading to inconsistencies and errors if not managed carefully.

Choosing the Right Approach for Your RAG Application

The choice between vector databases, knowledge graphs, or a hybrid approach for your RAG application ultimately depends on your specific use case and the nature of your data. Vector databases are often the best choice for getting started, offering ease of implementation and good performance for many applications. They're particularly well-suited for handling large volumes of unstructured data and performing semantic search.

Knowledge graphs, on the other hand, shine in scenarios that require a deep understanding of relationships and hierarchies within data. They're particularly valuable for recommendation systems, applications that need to recognize and utilise data hierarchies, and scenarios where explainability and clear data lineage are crucial.

Hybrid approaches can be beneficial when dealing with large, diverse datasets or when queries are complex and multifaceted, requiring both contextual understanding and detailed relationship mapping. However, the added power of hybrid systems comes at the cost of increased complexity and maintenance requirements.

As the field of RAG continues to evolve, we're likely to see further innovations in how these technologies are implemented and combined. The recent introduction of Microsoft's GraphRAG, for instance, highlights the growing importance of knowledge graphs in complex AI applications. As organisations continue to explore and refine these approaches, we can expect to see increasingly sophisticated RAG systems that can handle more complex queries and provide more accurate, contextual responses.

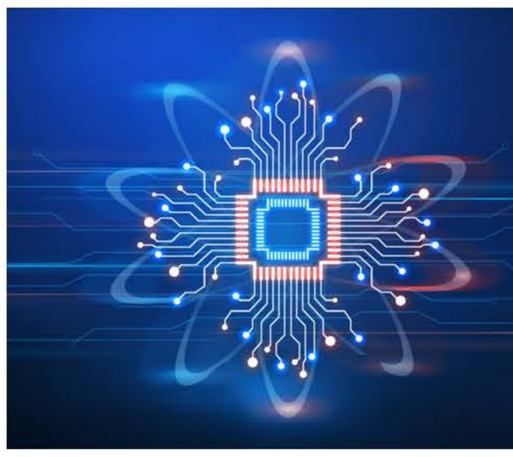
Conclusions

Large Language Models and Retrieval Augmented Generation represent a significant advancement in the field of natural language processing and information extraction. By addressing the limitations of traditional LLMs, RAG systems offer a powerful solution for enterprises requiring precise and reliable information extraction at scale. As the technology continues to evolve and more RAG offerings become available, we can expect to see increasingly sophisticated applications of this approach across various industries. The combination of vast language understanding capabilities with accurate information retrieval promises to unlock new possibilities in how we interact with and extract value from large-scale data repositories.

PROF. MANISH SHRIVASTAVA

IIIT HYDERABAD

is a professor at the Language Technologies Research Centre (LTRC). His research interests include machine translation and natural language processing for Indian languages. He is also the co-founder of Subtl.ai, a document retrieval platform that leverages GenAI to optimise information discovery on the internet.



Making ML Accessible, Collaborative and More Efficient With Google Cloud's Vertex AI

In order to make machine learning more accessible and useful for developers and businesses, Google created an end-to-end machine learning platform, known as the Vertex AI. The platform unifies Google Cloud's existing ML offerings into a single environment for efficiently building and managing the lifecycle of ML projects, thereby helping data scientists and ML engineers accelerate their ML experimentation and deployment. Here's a brief video shared by Google showing how this toolset supports ML workflows from data management all the way to predictions.

A presentation made by Google at the recently held TechForward Seminar series showcased the capabilities of Vertex AI. Some of the salient features that were discussed are as follows:

Vertex AI is a fully-managed, unified AI development platform for building and using generative AI. Developers and businesses can access and utilize Vertex AI Studio, Agent Builder, in addition to 150+ foundation models

Vertex AI is a machine learning (ML) platform that lets you train and deploy ML models and AI applications, and customise large language models (LLMs) for use in your AI-powered applications. Vertex AI combines data engineering, data science, and ML engineering workflows, enabling your teams to collaborate using a common toolset and scale your applications using the benefits of Google Cloud.

Vertex AI provides several options for model training and deployment:

- AutoML lets you train tabular, image, text, or video data without writing code or preparing data splits.
- Custom training gives you complete control over the training process, including using your preferred ML framework, writing your own training code, and choosing hyperparameter tuning options.
- Model Garden lets you discover, test, customise, and deploy Vertex AI and select open-source (OSS) models and assets.
- Generative AI gives you access to Google's large generative AI models for multiple modalities (text, code, images, speech). You can tune Google's LLMs to meet your needs, and then deploy them for use in your AI-powered applications.

After you deploy your models, you can use Vertex AI's end-to-end MLOps tools to automate and scale projects throughout the ML lifecycle. These MLOps tools are run on fully-managed infrastructure that you can customise based on your performance and budget needs.

A community initiative, anchored by

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

IIIT Hyderabad

TechForward
DISPATCH