modak

Beyond Buzzword:
The Power of Specialized GenAI

In the age of rapid technological advancement, the fascination with general-purpose language models (LLMs) has reached a fever pitch. These models, celebrated for their ability to generate human-like text, are often perceived as the pinnacle of AI (Artificial Intelligence) achievement. However, beneath the surface of this enthusiasm lies a stark reality: general-purpose LLMs frequently fall short when tasked with addressing the specific needs and stringent requirements of enterprises. While their broad capabilities are impressive, they often lack the precision, security, and cost-efficiency necessary for real-world applications. This whitepaper takes you into Modak's journey of developing specialized Generative AI (GenAI) solutions, highlighting the unique challenges we faced, the innovative strategies we employed, and the transformative impact of our work.

THE LIMITATIONS OF **GENERAL-PURPOSE LANGUAGE MODELS**

General-purpose language models, such as those based on the GPT-3 and GPT-4 architectures, have garnered significant attention for their ability to generate coherent text, answer questions, and perform various natural language processing (NLP) tasks. These models are trained on vast datasets, allowing them to perform well across a wide range of topics. However, when it comes to enterprise applications, they often encounter several critical roadblocks:

Cost:

Deploying and maintaining large, general-purpose models is an expensive endeavour. The computational resources required are substantial, leading to high operational costs. Furthermore, the need for continuous fine-tuning and updates to maintain accuracy and relevance adds to these expenses. For many enterprises, the cost associated with general-purpose models is simply not justifiable, especially when the performance does not meet their specific needs.

Data Security:

In the enterprise world, data security is paramount. Companies deal with sensitive information that must be protected from breaches and misuse. General purpose large language models are often available as APIs from LLM provider. Thus, for scenarios that involve sending enterprise data to the LLMs for RAG, fine tuning, or prompt engineering use cases, the data leaves enterprise network which may not be acceptable for sensitive data.

Accuracy:

While general-purpose models perform well on a wide range of tasks, their responses often lack the precision and contextual relevance needed for specific enterprise applications. This can lead to suboptimal performance and user dissatisfaction. For example, in customer service applications, a general-purpose model might generate responses that appear to be technically correct but fail to address the specific concerns of the customer or may not recognize the business context, leading to frustration and a poor user experience.

THE CASE FOR **SPECIALIZED GenAI**

To overcome these limitations, Modak advocates the use of specialized language models for GenAI solutions. By focusing on narrower, more defined tasks, specialized models can deliver higher accuracy, better performance, and improved cost-efficiency. This approach allows us to tailor the models to the unique requirements of each application, ensuring that they meet the specific needs of our clients.

To illustrate the practical application of specialized GenAI, we present a case study involving a global manufacturer of power tools. The customers of these power tools are independent contractors, who need to use these tools for their work. When they need to troubleshoot any issue, they would need to get on

a call with the support staff who would direct them to an expert to resolve their queries. Sometimes the part would need replacement and for that they would need to go to a store, with the uncertainty of the part not being available.

The client had an existing chatbot which relied on static decision trees and therefore had limited functionalities. The chatbot could answer only specific queries which relied on specific inputs from the user. Any deviation in the query asked or input provided would result in unsatisfactory response. Often the customer would end up calling the customer care for even simple queries. The result was prolonged query resolution times and decreased customer satisfaction.

THE **ASK**

The client sought an AI-driven customer support agent capable of providing quick, accurate responses to user TTinquiries. The agent should be able to handle wide variety of queries related to availability of parts and troubleshooting equipment. The agent should provide answers that are accurate and not stray from the knowledge base. Any PII inputs entered by the user should be identified and deleted, to comply with local regulations.

INITIAL APPROACH: **GENERAL-PURPOSE MODEL**

Our initial attempt involved creating a chatbot using a single large language model (LLM), experimenting with variants of ChatGPT. Despite initial progress, the performance fell short of expectations. The complexities of handling multiple functions (intent recognition, entity extraction, response generation, etc.) within a single model led to an overly complex design. The broad and generic nature of the model resulted in responses that often lacked the specific contextual relevance needed. Additionally, the resource demands were substantial, making the solution expensive to deploy and maintain. **Overall, the chatbot achieved only 80% success, far below the desired 99% accuracy.**

REVISED APPROACH: **SPECIALIZED MODELS**

Recognizing the limitations of our initial approach, we shifted to developing specialized LLMs. The specialized language models are trained to perform a narrow task. Being smaller models, a narrow task improves the odds for getting an accurate response. To identify the narrow tasks for the model, we created a customer support workflow which comprised different steps and identified areas where a specialized language model can perform well.

We identified two areas where specialized language model can perform well – detecting intent of the user (whether user is looking to enquire about availability of product or asking for help to fix a problem, or just saying hi) and detecting what product, model, part the user is referring to.

The workflow was designed to ensure that right path can be taken based on previous input. The workflow was flexible to accommodate multiple different questions in the same session.

For the first case, the specialized language model was trained to infer intent based on a user input.

For the second case, the specialized language models were trained with a named entity recognition approach to identify entities such as product name, a model number, or a part name in a user input.

Through iterations of training data, hyperparameter tuning, we were able to get more than 99% accuracy. An automated test would be performed to check accuracy after each iteration. We started with a 7b parameter model and then finetuned it for our use case.

The iterative process involved continuous assessment and optimization, including the development of robust test cases to evaluate model accuracy, improvement of tokenization and formatting of training data, and the use of quantization techniques to reduce model size and computational requirements without compromising performance. The result was a highly accurate, efficient chatbot system, achieving over 99% performance and ready for production deployment.

SPECIALIZED LANGUAGE MODELS: **THE ADVANTAGES**

The shift to specialized language models offers several significant advantages:

01 **Cost-Efficiency:**
Specialized models are cheaper and quicker to train, making them more cost-effective.

02 **Performance:**
By focusing on narrow tasks, these models deliver higher accuracy and better performance.

03 **Security:**
Specialized models can be localized, enhancing data security and compliance with regulations.

04 **Scalability**
Easier to scale and maintain, with shorter development cycles.

We achieved notable success with models like Dragon-Mistral (7B), gemma-2b-it, gemma-7b-it, and bling-phi-3-mini, demonstrating the potential of specialized LLMs in real-world applications.

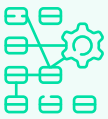
Implementing specialized LLMs requires robust infrastructure capable of provisioning on-demand compute resources. Fine-tuning involves adapting pre-trained models to specific tasks through further training, while inference uses pre-trained models to generate outputs based on given inputs. Deploying models on the cloud with the ability to scale up or down based on usage ensures cost control and performance optimization. Modak's approach abstracts infrastructure complexities, enabling seamless provisioning and s caling of resources.

USE CASES **AND APPLICATIONS**

Beyond customer support, specialized LLMs have broad applications across various industries. In the banking sector, they can streamline the management of voluminous operational manuals. In government, they enable citizens to query regulations in multiple languages through natural language interfaces. These applications highlight the versatility and effectiveness of specialized GenAI solutions, demonstrating their potential to transform enterprise operations.

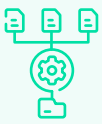
CHALLENGES AND **LESSONS LERNED**

The journey to specialized GenAI is not without its challenges. Creating a comprehensive knowledge base requires gathering and structuring vast amounts of information. Automating data structuring enhances performance, while optimizing embeddings improves information retrieval. Addressing response time issues with real-time generation, integrating audio for seamless interactions, ensuring clean and appropriate user inputs, and achieving smooth system integration are all critical challenges that must be navigated.



Knowledge Base Creation:

A robust knowledge base is essential for the accuracy and relevance of specialized GenAI models. This involves gathering comprehensive information from various sources and structuring it in a way that is accessible and usable by the AI. Automating this process can significantly enhance performance, reducing the time and effort required to maintain and update the knowledge base.



Data Structuring Automation:

Effective data structuring is crucial for the performance of specialized GenAI models. Automating this process ensures that data is consistently formatted and organized, enhancing the model's ability to retrieve and utilize information. This reduces the potential for errors and improves the overall accuracy of the AI.

Response Time Issues:

In real-time applications, response time is a critical factor. Ensuring that the AI can generate responses quickly and accurately is essential for maintaining user satisfaction. Addressing response time issues involves optimizing the model's processing capabilities and ensuring that it can handle high volumes of queries without compromising performance.

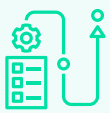
Profanity Checks:

Ensuring clean and appropriate user inputs is essential for maintaining a professional and respectful user experience. Implementing profanity checks involves developing external modules that can detect and filter out inappropriate language, ensuring that the AI interacts with users in a respectful and appropriate manner.

System Integration:

Achieving smooth integration of various components is critical for the seamless operation of specialized GenAI models. This involves ensuring that all parts of the system work together effectively, enabling the AI to retrieve and utilize information accurately and efficiently. Effective system integration reduces the potential for errors and enhances the overall performance of the AI.

KEY CONSIDERATIONS FOR **IMPLEMENTING SPECIALIZED GenAI**



Identifying the Right Use Cases :

The first step in implementing specialized GenAI is identifying the right use cases. Not all problems require a specialized approach, so it is important to evaluate where the highest impact can be achieved. Tasks that are complex, repetitive, or require a high degree of accuracy are prime candidates for specialized models.

Data Quality and Availability:

Specialized models rely heavily on the quality and relevance of the data they are trained on. Ensuring that you have access to high-quality, domain-specific data is crucial. This might involve data cleansing, enrichment, and integration from various sources to create a comprehensive and accurate training dataset.

Building the Right Team:

Implementing specialized GenAI requires a team with diverse skills, including data scientists, domain experts, AI engineers and User experience designers. Collaboration between these groups is essential to ensure that the models are not only technically sound but also aligned with business objectives.

Embracing a Culture of Innovation:

For enterprises to fully harness the power of specialized GenAI, embracing a culture of innovation is crucial. This involves fostering an environment where experimentation is encouraged, and failure is seen as a learning opportunity. By continuously exploring innovative ideas and pushing the boundaries of what AI can achieve, organizations can stay ahead of the curve and drive continuous improvement.

Ensuring Ethical AI Practices:

With the increasing use of AI comes the responsibility to ensure that it is used ethically. This includes addressing issues such as bias, transparency, and accountability. By adopting robust ethical frameworks and governance structures, organizations can build trust and ensure that their AI solutions are fair, transparent, and aligned with societal values.

SPECIALIZED GenAI:
A PATHWAY TO 99% ACCURACY

Through our work, we have observed a recurring issue with general-purpose LLMs: many clients find that their models only achieve around 85% accuracy in specialized enterprise use cases. This level of accuracy is often insufficient for applications requiring high precision. As a result, the output falls short of business expectations, leading to frustrations and the need for more advanced solutions. Modak's specialized approach to GenAI has consistently achieved much higher accuracy rates, often reaching 99% or more. This is particularly important in complex tasks such as querying graph models or managing intricate workflows where even minor errors can have significant consequences. By tailoring models to specific domains, we ensure that our solutions not only meet but exceed the expectations of our clients. Furthermore, our ongoing work in developing models capable of effectively interfacing with graph databases underscores our commitment to pushing the boundaries of what AI can achieve.

The journey from general-purpose to specialized GenAI is a transformative one, offering enterprises the o pportunity to harness the full potential of AI for real-world applications. By focusing on specialized models, Modak has overcome the limitations of generic solutions, delivering higher accuracy, better performance, and improved cost-efficiency. As the field continues to evolve, the lessons learned, and approaches developed in our journey will pave the way for future innovations in specialized GenAI.

modak

AWARDS AND **RECOGNITIONS**

01 Gartner

- Featured in the 2024 Gartner® Market Guide for DataOps Tools
- Cited in the Gartner® 2023 Report on Data-Centric AI Solutions for Streamlining AI Development
- Recognized in the Gartner® 2022 Market Guide for Active Metadata Management
- Highlighted in the 2021 Market Guide for Data & Analytics Governance Platforms

02 Bio-IT World

- Best of Show Award - Data Platform Innovation, 2022

03 Cloudera

- Global ISV Partner of the Year, 2022

04 Strata - Data Conference

- Data Platform built in record time, 2021

Ready to transform your data into a strategic asset?
Contact Modak today for a free consultation and let's discuss your path to data dominance.

CONTACT US

- **Find out more**
  https://modak.com/contact/
- **Follow us**
  www.linkedin.com/company/modak/

USA
21660 W Field Parkway, Deer Park, IL – 60010

USA
312 S, 4th St, Suite 700, Louisville, KY 40202

UAE
Dubai Silicon Oasis (DSO), 341041

INDIA
Elemental #337, Malakunta, Financial District, Nanakramguda, Telangana 500032