Customising generative AI: Harnessing document retrieval and fine-tuning alternatives for dynamic marketing insights

Received (in revised form): 12th March, 2024

Dakota Crisp
Senior Manager of Data Science, OneMagnify, USA

Dakota Crisp is a senior manager of data science at OneMagnify. With a PhD from the University of Michigan, his hypothesis-driven approach to investigating complex systems provides a distinctive take on consumer behaviours. Dakota is an exceptionally skilled writer who leads many of OneMagnify's internal and external publications. His passion for leadership enables his team to transform marketing analytics into innovative solutions.

OneMagnify, 12N Main St #100, Ann Arbor, MI 48104, USA
Tel: 001 888 294 1512; E-mail: dcrisp@onemagnify.com

Jacob Newsted
Data Engineer and Data Scientist, OneMagnify, USA

Jacob Newsted, a data engineer and data scientist at OneMagnify, boasts a multifaceted background in computer science. His expertise spans an impressive array of disciplines, including game development, web development, data engineering and data science. This diversity enables him to approach various problem spaces with unique perspectives. Jacob earned his MS in machine learning and evolutionary computation from Michigan State University. He is passionately committed to continuous learning and applying cutting-edge advancements in these fields to his work.

OneMagnify, 12N Main St #100, Ann Arbor, MI 48104, USA
Tel: 001 888 294 1512; E-mail: jnewsted@onemagnify.com

Brendon Kirouac
Data Scientist, OneMagnify, USA

Brendon Kirouac is a data scientist at OneMagnify. He earned his BS in physics from Wayne State University and shortly after began work in L4 autonomous vehicle development as a systems and integration test engineer focused on motion planning and controls. Observing the ability of machine learning to enable a vehicle to perceive and navigate the world inspired him to pursue data science. At OneMagnify, Brendon is working on designing and implementing consumer-facing end-to-end generative AI solutions.

OneMagnify, 12N Main St #100, Ann Arbor, MI 48104, USA
Tel: 001 888 294 1512; E-mail: bkirouac@onemagnify.com

Danielle Barnes
Senior Director of Data Science, OneMagnify, USA

Danielle Barnes is a senior director of data science at OneMagnify. She is an accomplished analytics leader with extensive experience across the entire data life cycle. Her work directing enterprise analytics initiatives for companies across various industries has made her a powerhouse at realising visions in complex environments. She is a Spartan superfan with a BA in mathematics and an MS in statistics, and she is currently pursuing a PhD in data science, all from Michigan State University.

OneMagnify, 12N Main St #100, Ann Arbor, MI 48104, USA
Tel: 001 888 294 1512; E-mail: dbarnes@onemagnify.com

Catherine Hayes
Senior Director of IT, OneMagnify, USA

Catherine Hayes is a senior director of IT at OneMagnify. Her experiences as a full-stack engineer, small business owner, teacher and technology leader have contributed to her exceptional ability to help break down abstract technical concepts and translate client requirements into functional and elegant solutions. She leads a team of data engineers, software developers and solution architects.

OneMagnify, 12N Main St #100, Ann Arbor, MI 48104, USA
Tel: 001 888 294 1512; E-mail: chayes@onemagnify.com

Jonathan Prantner
Chief Analytics Officer, OneMagnify, USA

Jonathan Prantner is the Chief Analytics Officer at OneMagnify. His approach to applied mathematics has pushed analytics to the limits for over two decades. Jonathan's career has spanned educational research, automotive, consumer packaged goods, travel and healthcare. At OneMagnify, he leads efforts surrounding applied artificial intelligence and machine learning as well as integrating advanced analytics with data visualisation platforms. Jonathan is a celebrated thought leader and the recipient of multiple data science patents.

OneMagnify, 12N Main St #100, Ann Arbor, MI 48104, USA
Tel: 001 888 294 1512; E-mail: jprantner@onemagnify.com

Abstract This study delves into the transformative impact of leveraging large language models (LLMs) in marketing analytics, particularly emphasising a paradigm shift from fine-tuning models to the strategic application of document retrieval techniques and more. Focusing on innovative methods, such as retrieval augmented generation and low-rank adaptation, the paper explores how marketers can now activate against vast and unstructured datasets, such as call centre transcripts, unlocking valuable insights that were previously overlooked. By harnessing the power of document retrieval and adaptation, marketers can bring their data to life, enabling a more nuanced and adaptive approach to understanding consumer behaviour and preferences. This research contributes to the evolving landscape of applied marketing analytics by demonstrating the efficacy of document retrieval in enhancing the utilisation of LLMs for dynamic and data-driven marketing strategies.

KEYWORDS: generative AI, marketing analytics, call centre, natural language processing, document retrieval techniques, retrieval augmented generation, low-rank adaptation

INTRODUCTION

Artificial intelligence (AI) is a broadly used term in marketing today. Many systems and solutions advertise the fact that they are powered by AI, but AI can mean many things to many people. That is why it is important to outline a common definition for AI when considering how to apply it to marketing use cases. In its simplest definition, AI is a system that makes decisions based on data and then acts upon them. Equally important to its definition is the fact that AI attempts to mimic an existing human process.

The evolution of AI has been a journey, progressing from the foundational principles of machine learning to the sophisticated capabilities of modern large language models (LLMs). Machine learning, at its core, seeks to mimic the process of differentiation. By using a set of quantifiable attributes to identify an object class or to estimate a

specific value, machine learning models can replicate the way in which people categorise and estimate based on measurable features.

In contrast, generative AI mimics the nuanced human experience of sensory familiarity, which, conceptually, is not as concrete of a process as those undertaken by machine learning. Sensory familiarity is the cognitive process of recognising and feeling a sense of comfort or recognition in response to sensory cues. These cues can be visual, auditory or tactile. When people encounter a new environment, object or experience, they intrinsically compare it to past experiences. The extent to which it resonates with positive past sensory memories creates a feeling of familiarity and ease. This is a highly subjective process that relies on the brain's ability to associate current sensory input with stored memories.[1]

In the context of AI, particularly in generative models like LLMs, the concept of sensory familiarity is simulated by using keywords to link prompts to relevant information within the model, allowing the AI to provide responses that align with the user's query in a manner reminiscent of human associative thinking. This allows LLMs to link the user's input back to the appropriate information within the model, providing a response that aligns with the query. This is similar to the human process of associating current sensory experiences with memories. By mimicking more complex and intricate human processes, LLMs create opportunities for advancement in many fields.

While publicly available LLMs have a vast amount of information within their training data, one of the biggest advances has come from marketing analytics professionals leveraging these tools to better interact with and gain advantages from their own proprietary data. Companies leveraging their own data is by no means new, but a large barrier in the past was that, despite being rich with information, many data sources were simply too unstructured to be used. LLMs have been used to sort through and derive meaning from unstructured data with an efficiency that humans cannot rival. In instances where the LLM has issues with specific tasks or domains, researchers have implemented fine-tuning methods.

Fine-tuning is the process of continuing to train a general, pretrained model on a small subset of specific data. The goal of this process is to preserve the LLM's general language knowledge while also imparting new knowledge and context from unique and/or proprietary datasets.

Fine-tuning is easy to implement, as it typically uses the same procedure to train as the original pretrained model (text completion for GPT, masked language modelling for bidirectional encoder representations from transformers, etc), and only the single model is used. However, the major issue with fine-tuning is that fine-tuned models can sometimes lose catastrophic amounts of knowledge acquired from the initial pretraining. For example, fine tuning Llama V2 7B on a subset of the alpaca-52k dataset has been shown to improve mathematics capability by 8 per cent, but causes degradation in all other evaluated task types, like procedure, reasoning and writing.[2] Thus, the difficulty with fine-tuning is gauging exactly what data to fine-tune on to get the most benefit, without corrupting the initial, pretrained model.

Unfortunately, there is no foolproof method that balances this correctly as it largely relies on experimentation for each specific use case. So, while improving a model for a specific task, the model's capability for anything else is degrading. This makes fine-tuning undesirable for any implementation where additional functionality or use cases will be added over time, which is a lot of use cases. Fortunately, there are alternatives to fine-tuning that avoid modifying the foundational model while still achieving improvements for target use cases.

This paper provides a high-level overview of two alternatives to fine-tuning, discusses

how marketing analytics professionals can utilise these techniques to leverage their data and highlights two case studies where these techniques could be successfully implemented.

DOCUMENT-POWERED GENERATIVE AI

Explanation of retrieval augmented generation and low-rank adaptation methodologies

Retrieval augmented generation

Retrieval augmented generation (RAG) is a technique developed to impart new information to a foundational model without requiring the difficult process of fine-tuning.[3] The idea behind RAG is to use an auxiliary model to extract the *meaning* behind a piece of text and then select a small subset of the most relevant knowledge to retrieve from a dataset based on how similar it is. The semantic meaning of a piece of text would be represented by a list of numbers (a semantic vector) which could then be compared to another text's representative list of numbers to find how similar they are. Figure 1 shows, on the left, how each document is transformed with an encoder model, creating a vector that represents the meaning of the given text. Then it is paired with its original raw text and put into a database for retrieval when the user asks a question. On the right, the figure shows the user asking a question, which is encoded by the same encoding model as on the left. The vector is then compared with every vector representation of the data in the database, and only the most similar documents are returned. The LLM is left out, as it is not specifically part of the RAG system but rather used to enhance the output of RAG systems. For example, the sentences 'I love cats' and 'I love kittens' should have very similar vectors. However, both sentences would be vastly different from the sentence 'The first law of thermodynamics describes the conservation of energy'.

This technique enables analytics professionals to select only the most relevant data within their own dataset to contextually modify any query, as opposed to needing to fine-tune the model or inject the entirety of their dataset into every query. This methodology requires no training or fine-tuning, making it quick and easy to implement. Furthermore, this methodology is designed to be general-purpose, and is sufficient for many common problems. Of course, language is difficult and nuanced. RAG models find general patterns in language, which can sometimes lead to spurious conclusions. They can also have issues with dialogue because the text is not structured in a standard format.

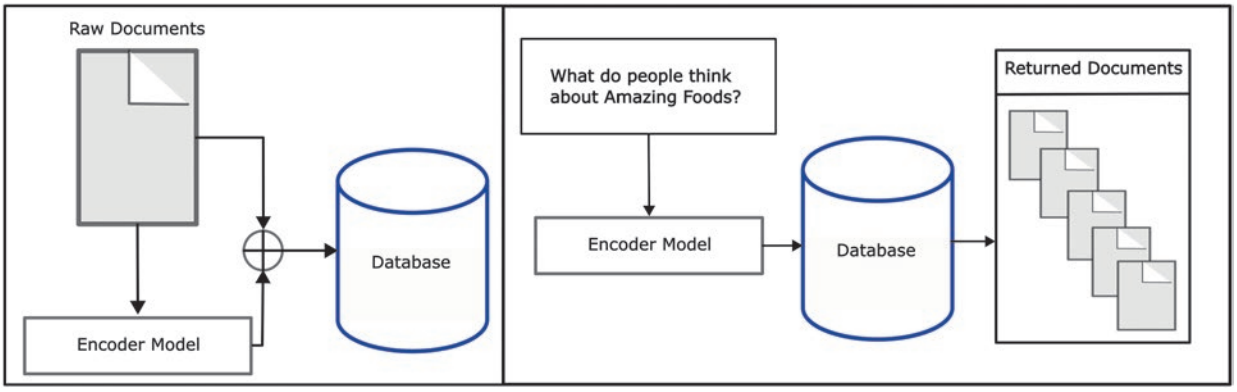

Figure 1: RAG system

Low-rank adaptation

Using adapters is another method of imparting new knowledge onto an existing LLM. An adapter is an algorithm that modifies the inputs, inner parameters and/or outputs of an LLM, thereby modifying the model without directly changing all its parameters. Low-rank adaptation (LoRA) is a popular implementation that trains a surrogate set of model parameters completely separately from the original model.[4] Figure 2 shows, on the left, that when a model is fine-tuned under standard conditions, all parameters in the model are updated. On the right, the figure shows that when fine-tuning using LoRA, the main model's parameters are frozen and do not update. Instead, a new set of parameters is updated (on the right) and only that smaller set of parameters is updated.

The adapter does not change the original model and usually has far fewer parameters (as small as but not limited to 0.01 per cent of the initial parameters), which means far fewer resources are necessary to fine-tune the new model and there is no chance of corrupting the foundational model. The trade-off for this safer implementation is an increase in complexity and a decrease in accuracy. Furthermore, because this method utilises a separate set of parameters, it is possible to employ multiple LoRAs at the same time.[5] A framework for hosting concurrent LoRAs is S-LoRA, which stores each adapter in CPU memory, and pulls the currently used adapters into GPU memory when needed. The comparatively small size of LoRA adapters allows for this transfer to happen quickly.[6] A challenge with implementing these recent developments is incorporating them within existing libraries or new models. Sometimes, a new model architecture will release, and it can be weeks or months until it is supported by open-source libraries. With a more bare-bones structure, the amount of dependencies, and by extension failure points, decreases.

Illustration of how these methods empower marketers to activate against extensive and unstructured datasets

RAG systems are designed to *search* through large sets of data using natural language. The value that RAG systems add to data ecosystems is that analytics professionals no longer need to use SQL, fuzzy string matching or other complicated

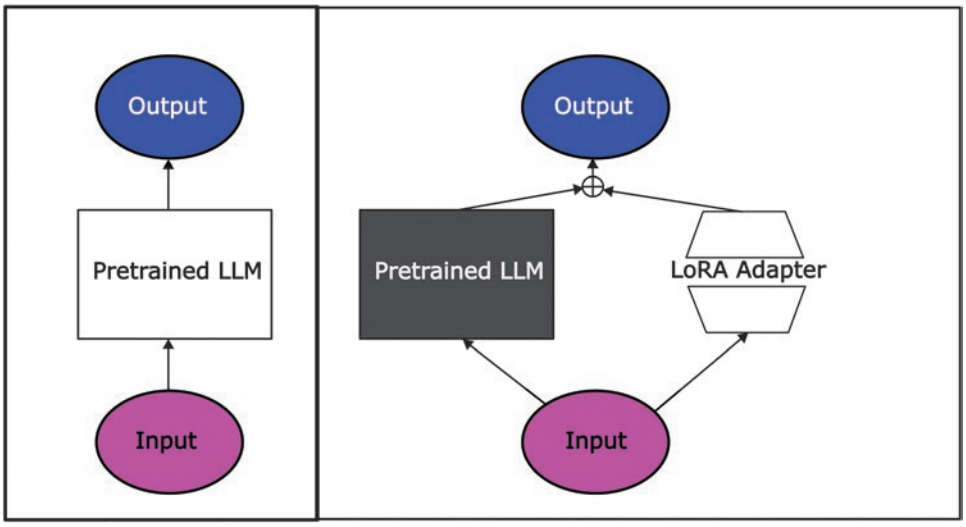

Figure 2: LoRA model

programming techniques to explore data. Instead, they can use *their* voice and words to find relevant information.

While RAG systems are similar in architecture between every implementation, the data decides each component's inner workings. Figure 1 outlines the general flow of a RAG system, and the following sections explain each component in detail and how it changes depending on the data.

Embedding models

The embedding model is what transforms the data into something that a machine can then relate to user questions and queries. While it is tempting to look at benchmarks to see which embedding model is the most performant, all models are not built equally, and benchmarks are only the surface-level deciding factor. There are three general flavours of embedding models: general, instruction fine-tuned and specialised.

General models perform well on most types of data but do not excel with any one type. AnyLM, GTE and BGE are examples of general models and are all trained on a plethora of text from many sources (papers, articles, social media, etc). These models should be considered for general use cases (question and answer, document search, sentiment analysis, etc) or as a starting point for new data scientists.

Instruction fine-tuned models take the idea of using a specific *voice* to query data to the next level. Rather than using large amounts of unstructured data to train the initial model, a more natural language structure (in the same way a person might ask another person a question) is used to transform the data into vectors. This is especially strong for companies that see their user base trying to query their data with more conversational questions. Examples of these models are E5-Instruct and Instructor.

There are also highly specialised types of models that are trained on data from specific sources, such as PubMed or Arxiv. This specificity limits the general usefulness of these models but makes them exceedingly powerful in cases where the queried data matches the training source. These models tend to be fine-tunings of existing general models and can be found on open-source model hubs such as HuggingFace.

Foundational LLM

The final part of the RAG LLM chain is the LLM itself. One major consideration at this step is the sensitivity and privacy of data. OpenAI's ChatGPT/GPT4, Google's Gemini and other proprietary models are exceedingly strong. However, they can also have very relaxed data privacy standards. If privacy is an issue, then consider using self-hosted, open-source models where data can be kept in an isolated ecosystem.

LoRA and its application to voice

Foundational language models are trained on a large corpus of data — commonly over trillions of tokens that cover a variety of topics and sources. This volume and variability of data enable them to excel in general usage but impede these models from developing a specific voice. Why does this matter? In the context of a company that generates content for multiple brands, staying true to a brand's identity in all content is crucial. Using LoRA, a language model can be fine-tuned on a dataset that is orders of magnitude smaller than the initial training data without modifying the underlying model. Improvements can be seen from fine-tuning on as few as 50 data points, with significant improvements realised for about 1,000 data points, dependent on the length of the data.[7] The non-destructive nature of LoRA makes it possible to create an infinite number of fine-tunes on a foundation model in a memory-efficient manner. For a solution involving different brands or voices, different LoRA fine-tunes can be used on

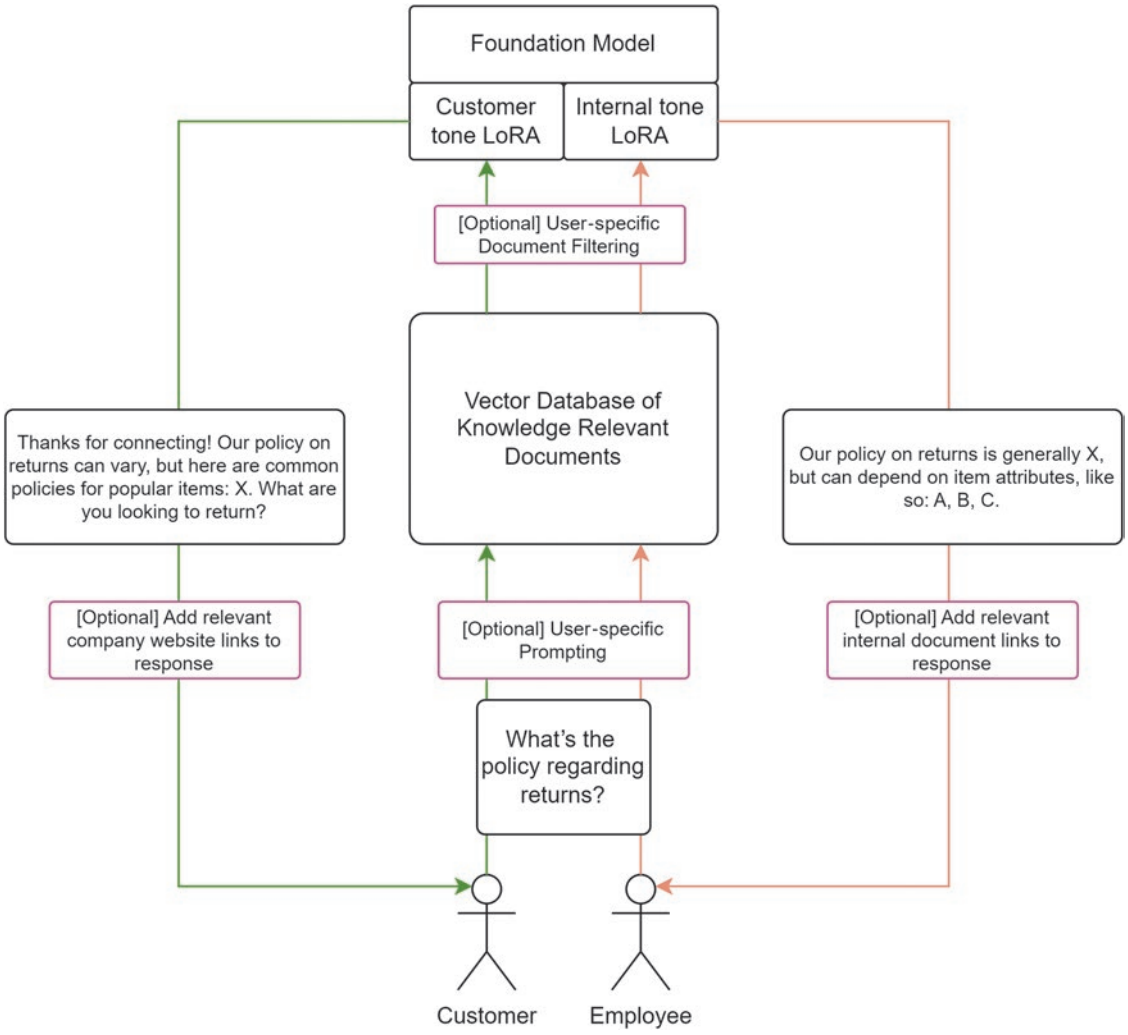

Figure 3: How LoRAs fine-tuned for different tones can complement a RAG solution

the same base model easily without needing to replicate the foundation model weights multiple times (see Figure 3).

LoRA can also be useful for creating content across different media channels. The way companies write content for LinkedIn versus Instagram can vary, and it could be unsettling for customers to see a LinkedIn-style post on Instagram. If companies have access to an archive of posts for different media, LoRA can put that data to use and create media-specific model fine-tunes. Similarly, there are differences in the way a business communicates internally versus directly to customers. If text samples can be collected for each audience, it becomes possible to use LoRA to create an LLM geared towards generating text appropriate for internal and external communications.

Combining the strengths of LoRA with RAG, it is possible to build solutions where multiple LoRA models have access to the same pool of use-case-specific data. Work would not need to be duplicated to augment each model with knowledge, though it could be if desired. Most RAG solutions allow individuals to combine hard filtering with semantic similarity searching. If a dataset has

associated metadata that determines whether it is private or public, for example, it is trivial to only allow certain workflows access to certain data. This way, a LoRA fine-tuned model used for internal communication can have access to private data, but one meant for customer communications can be blocked from such data.

REVOLUTIONISING AUTO DEALER CUSTOMER REVIEWS

Review data is a powerful source of customer feedback that can provide insights into a brand or product's strengths and shortcomings. One such example of this is from automotive dealerships and the aggregate review websites that store customer reviews. Once the data is scraped, cleaned and stored into some sort of raw text format (CSV, TXT, database, etc), a RAG system can be used to enhance understanding of this large corpus of text.

For this use case, a RAG-empowered chatbot is desirable to act as the middleman between a user (such as a dealership owner or a car manufacturer) and the rich dataset of dealership reviews. A user might interact with a standard chatbot UI to ask it questions such as 'What are users saying about Honda Civics at Great Dealership?'. Leveraging RAG, the chatbot could then return an answer backed by real reviews and not just general knowledge from the LLM. Providing a ground truth for the LLM to work from reduces instances where answers are not based in reality. Using RAG for the purpose of generating trivia questions and answers has been shown to achieve 43 per cent higher factual accuracy than the same generative LLM without a retrieval component.[8] In practice, the RAG system first encodes the question the user asks into a list of numbers, or a vector, to compare against the vectors of each respective review in the dataset. It matches the most relevant reviews to the question before sending them to the LLM and asking it to summarise the output into an easy-to-read, concise description.

From the standpoint of grammar and syntax, reviews can be messy and sometimes need additional techniques to more easily find the most relevant information. One such method is to use an LLM to reword the question many times and then encode all the permutations. The new vectors can then be used to get many more unique reviews. Alternatively, because of how output semantic vectors function, the average of the question vectors can be used to gain a 'middle-ground' question that might represent a more general form of what the user is trying to ask.[9]

Additionally, sometimes analysts may want to have more control over what is returned by the RAG system. In instances like these, it is possible to filter the documents that RAG can return. For instance, if a user of this chatbot only wants to analyse reviews from a certain source (such as a specific website), mechanisms can be implemented that filter out reviews from any other source before using the RAG system. Given that some sources (such as social media and Reddit) have different styles, formats and levels of professionalism, this can have a large impact on the kind of insights derived.

UNLEASHING THE POWER OF CALL CENTRE TRANSCRIPTS

Blog posts, articles, papers, reviews and social media posts are common sources of training data for traditional sentiment evaluation. While varied in topic and content, these are all unidirectional forms of information, meaning they represent a single perspective. Call centre data provides an interesting use case for RAG and fine-tuned systems because it represents two different perspectives. Recent advances in multi-source and multi-domain sentiment analysis have expanded the foundation and context by which context is understood, but do not

completely address the specifics of the call centre experience.[10]

Call centre transcripts are unique because at each turn in the conversation, ie the switching of the speaker role from agent to customer, the tone and trajectory of the discussion have the potential to change. In this regard, call centre transcripts can be viewed as a dance between agent and customer, and they require a specialised approach which blends traditional natural language processing (NLP) with the power of LLMs.

The following process is proposed for appropriately analysing call centre data:

- preprocessing;
- traditional sentiment scoring;
- emotional sentiment scoring;
- conversation trajectory; and
- NLP–RAG synthesis.

Preprocessing

Preprocessing is a crucial step in preparing transcripts for sentiment scoring. Beginning with text cleaning, where irrelevant elements such as special characters, punctuation and formatting are removed, each step must be customised to the form of call centre data that is being analysed. This step ensures that the transcript data is in a standardised and readable format while maintaining indications of which party is speaking and the sequence of events. References to the speaker and the order should be removed from the text fields and recorded in a separate linked variable, facilitating accurate sentiment analysis.

Tokenisation is another essential preprocessing step that involves breaking down the transcript into smaller units, such as utterances or sentences, known as tokens. Tokenisation enables a more granular analysis of the text, allowing sentiment analysis algorithms to operate at the sentence level. This process aids in capturing the nuanced structure of language, which is crucial for accurately discerning sentiment. Additionally, stemming and lemmatisation normalise words by reducing them to their root form, decreasing the dimensionality of the data and ensuring consistency.

Handling stop words is a critical aspect of transcript data preprocessing. Stop words, which are common words like ‘the’, ‘and’ or ‘is’, may not contribute much to the sentiment of a sentence and can be removed to focus on more meaningful content. However, for call centre data in particular, the careful selection of stop words is necessary since the removal of essential words might lead to a loss of context. This is where the use of the rapid automatic keyword extraction (RAKE) algorithm to understand the specifics of language within the call centre setting is helpful. RAKE is an algorithm used in NLP to extract keywords or key phrases from a text document. RAKE is a rule-based algorithm designed to be fast, simple and language-agnostic.

The RAKE algorithm begins by separating the text into n-grams, phrases of length n, providing a list of candidate keywords. Each keyword is assigned a score based on its complexity, ie the order of n in the n-gram and the frequency with which it appears in the document. Since the score is calculated by considering both the number of times an n-gram appears and the number of words it contains, longer phrases that appear more frequently are given higher scores.

There are inherent benefits to RAKE as well as some limitations. RAKE is effective for extracting keywords from large bodies of text where a quick extraction of recurrent themes is required. It is important to keep in mind that RAKE may not capture more complex relationships between words or consider the meaning of phrases, since it relies on statistical patterns of word co-occurrence and frequency.[11]

In the case of preprocessing call centre data, RAKE proves helpful in handling stop words. By highlighting frequently recurring themes and phrases within the transcripts,

utterances, such as greetings, standardised scripted introductions and general conversational transitions, can be removed from the text. This allows the removal of utterances such as 'yes', 'hi', 'thank you', 'how can I help you?', 'what seems to be the problem?' or 'have a good day' from being considered when looking at the sentiment of the discussion.

Handling negations is also important with spoken text to avoid potential misinterpretations. Techniques such as double negation resolution provide a more accurate representation of the emotional tone in the transcript data.

Traditional sentiment scoring

Sentiment scoring of transcripts analyses individual sentences within the text to determine the emotional tone expressed. This approach utilises algorithms to classify sentences as positive, negative or neutral, based on predefined sentiment lexicons or machine learning models.

One issue that often arises from this approach is the complexity of spoken language. Sentiment is context-dependent, and sentences may have different meanings when considered in isolation compared to the broader context of a conversation. While traditional NLP struggles to grasp subtle shifts in sentiment, it provides a solid basis of data points to be considered within a more holistic understanding of the interaction. Incorporating sentence-level sentiment as an input in future summaries helps address this issue.

Emotional sentiment scoring

The overall sentiment of a text is not always straightforwardly reflected in the aggregation of sentence-level sentiments. This is due to the dynamic nature of emotion in conversation, where the sentiment in one sentence may offset or dilute the emotional impact of another. Analysing sentiment at the sentence level without considering the potential cancellation effects can lead to a loss of emotional depth. This illustrates the importance of adopting sophisticated sentiment scoring approaches for conversational data.

The max delta aggregation method provides a more accurate representation of the complex emotional variation throughout the conversation. Instead of assigning equal weight to each sentence, this method seeks to capture the most significant changes in emotion throughout the text. Each sentence is assigned a score reflecting the emotional intensity, with higher values indicating more emotionally charged sentences, generally ranging from 0.2 to 0.9. These sentence-level scores are then aggregated to either a total utterance or total conversation level. The max delta aggregation identifies the largest change (the 'max delta') in each emotion between sequential sentences. These max delta values are then added together for all emotions or specific subsets, such as positive or negative emotions. This process filters out monotone portions of the conversation by giving lower scores to portions that consistently convey a single emotion and amplifies emotional variety and fluctuations in emotional intensity.

Conversation trajectory

The overall direction and outcome of the conversation can be measured through a concept we have termed 'emotional progression'. This allows the understanding of not only the sentiment portions of the conversation, but also how the agent is able to redirect the tone of the conversation. Combining exponential smoothing of the sentence-level sentiment with the max delta approach provides a framework for assessing the emotional progression of call centre transcripts. Exponential smoothing is employed to give greater weight to recent sentences while retaining the impact of the earlier portions of the discussion. This

enables a dynamic representation of the evolving emotional tone throughout the conversation. This smoothing technique combined with the removal of conversational stop words mitigates the impact of individual outliers and provides a more continuous and responsive measure of sentiment changes over time.

The max delta approach pinpoints the most significant emotional shifts within the call centre transcript. By identifying the largest change (max delta) in sentiment between consecutive sentences, this method captures pivotal moments in the conversation that contribute significantly to the overall emotional trajectory. Applying max delta to subsets of emotions, such as positive or negative sentiments, facilitates a fuller understanding of emotional dynamics and their impact on the customer experience.

The combination of exponential smoothing and the max delta approach yields a comprehensive metric for 'emotional progression' in call centre transcripts. This metric can be utilised to assess how effectively the agent addressed the customer's issues, by examining changes in sentiment from the beginning to the end of the conversation. A positive emotional progression may indicate successful issue resolution or improved customer satisfaction, while a negative progression may highlight areas that require attention or further training for agents.

This combined approach allows for the identification of the specific portions of the agent's responses that influenced the customer's experience. By analysing the sentences corresponding to the peaks and troughs in emotional progression, businesses can gain insights into the effectiveness of certain communication strategies or language choices employed by their agents. This level of granularity enables targeted improvements in agent training programmes, fostering a more emotionally attuned and customer-centric approach within the call centre environment.

NLP–RAG synthesis

By following the aforementioned steps, there is now a wealth of metadata associated with each utterance within each conversation in the call centre logs. Though useful, this exacerbates a common issue with call centre transcripts: the size of the underlying data. To mitigate this issue, daily batch processing of call centre transcripts and their metadata is used, producing more actionable summaries.

In a variation of the standard RAG approach outlined above, the metadata derived from each utterance can be passed to the LLM through an enhanced prompting technique. Figure 3 shows how LoRAs fine-tuned for different tones can complement a RAG solution. The same knowledge database can be used, but the way the information is conveyed to the end user can change substantially. Additionally, different behind-the-scenes prompts can be used for different users to further change tone and optional document filtering for different end users. For example, a metadata tag in your dataset can determine if data is 'confidential' and specifically not use confidential data for customers.

A sample prompt that directs the LLM to utilise its vast understanding of language to extract consistent insights across call centre transcripts:

> Summarise the following transcript into the five main themes of the conversation, classify the underlying topic based on the RAKE themes provided, include the overall sentiment as well as the emotional progression in each phase and include a summary of the agent and customers dispositions at the end of the call {call transcript with meta data}

At a base level, call centre data provides various conversations that can be used by agents in the future to resolve similar problems and reference prior pain points for customers. This relies on the ability of LLMs

Call summary	Theme	Subtheme	Sentiment	Emotional progression
Customer called to enquire about the unavailability of a heated steering wheel feature for a full-size truck.	Customer enquiry	Feature availability **Heated steering wheel** Full-size truck	0.1	0.1
Dealerships were unable to provide a reason for the unavailability.	Dealership communication	Inconsistency of information Unanswered question	-0.3	-0.3
Agent was unable to provide any information on why the feature was unavailable or if it would become available in the future.	Limited information		-0.5	-0.8
Customer expressed disappointment and may not purchase the vehicle due to missing features.	Customer dissatisfaction	Purchase decision **Lost sale**	-0.8	-1.2
Agent suggested looking at other models, like a sedan.	Alternative options	Sedan	0.4	-0.7
Customer frustrated by the fact that the suggestion is a different type of vehicle.	Customer dissatisfaction	Loss of trust	-0.3	-1.1
		Overall	-0.5	-1.1

OUTCOMES	
Customer action	Explore other makes and models
Agent action	Close case

Figure 4: A sample output from the standardised prompt

to properly mine the data and extract insights that are both appropriate and useful (Figure 4).

This could be done by combining RAG and LoRA to determine the areas of emphasis and relative weights across raw transcripts. For example, a RAG system could systematically archive prior call data in an accessible format. Once enough conversation summaries are encoded into the database, a future agent consulting a customer with a problem could automatically be shown the top three most relevant past conversations that were resolved successfully. Semantic similarity (matching how similar the current topic is to past topics) could be used to make sure the retrieved conversation logs were relevant. Furthermore, by filtering on the outcome of the conversation, the retrieved conversation logs would have an increased likelihood of leading to the best outcome for the customer.

To take things a step further, successful calls (such as successfully diffusing a problematic situation) can be the basis for training a LoRA to generate example phrases a customer service agent might use to better serve a current customer. The same structure of the stored data can additionally help create an automated chatbot that can be used as a first line of help for customers before reaching a human agent. Given prior conversations with a real customer service agent, a business could find conversations that started either well or badly but eventually resulted in a good ending. Fine-tuning an existing LLM like Llama-V2-Chat using LoRA over this kind of data is both an inexpensive and fast way to personalise a chatbot to more effectively deal with customers before it is necessary to transfer a customer to a real agent.

BRINGING DATA TO LIFE

Synthesis of findings from case studies

Using a technique like LoRA allows individuals to add additional parameters to a model that are specifically trained on their data. This is a great method for applications like changing the tone or style of responses from an LLM, as well as controlling output formatting. If the output of a foundational LLM is lacking in the style or voice that is needed, a fine-tune using LoRA and a small sample of data that contains the desired verbiage and style of writing can help in a non-destructive and parameter-efficient way.

Using a technique like RAG allows context to be dynamically provided to inform a response based on the input without modifying the base model whatsoever. This is most advantageous when trying to interact with and generate insights from unstructured domain-specific or proprietary data, enabling the creation of new data streams as well as eliminating inefficient workflows. For example, instead of manually reading through thousands of documents, a RAG system can provide relevant documents based on user prompts.

Highlighting the broader implications for marketers in activating untapped data sources

The general ability of generative AI to unlock new insights from unstructured data provides immense value for marketers. Creating a task-specific model using LoRA and/or RAG to utilise domain-specific data will provide an edge that cannot be obtained with a foundation model alone. Teams without the resources to train a generative model from scratch are now enabled to inexpensively create AI-powered solutions tailored to their goals. Used standalone or in tandem with other AI tools, like emotional sentiment prediction, valuable information can be extracted from otherwise inaccessible data sources. There is suddenly immediate value in data that has been kept dormant for years (call logs, reviews, internal documents, etc).

CONCLUSION

Recap of the shift towards document retrieval in LLM-based marketing analytics

Document retrieval methods are growing in popularity due to the quick implementation time and independence of the foundation model. Foundation models are being released by technology leaders in the generative AI space regularly, and that is not slowing down. There could be a model released next week that eclipses the current models for general-purpose tasks. Even if a company already has a RAG solution in place, the underlying model can be swapped out quickly to produce quality output changes. This model-agnostic property makes document retrieval a tool that will remain relevant as AI continues to develop.

Future directions and potential advances in the field

Given recent advances in generative AI, significant progress is expected to continue in the coming decade. Generative models that create and interpret charts (such as ChatGPT-4 Vision) are already available, so further applications for marketing analytics could include automated optimisation and report generation, which could enable quicker times to insights by integrating generative AI into widely used and accepted methods. Coupling this with content generation can provide a full suite of applications to customise content, effectively targeting consumers with a data-driven strategy to increase conversions and subsequent revenue.

References

1. Helmefalk, M. (2017) 'Multi-sensory Cues in Interplay and Congruency in a Retail Story Context: Consumer Emotions and Purchase Behaviors', Linnaeus University Press, Kalmar.
2. Ghosh, S., Evuru, C. K. R., Kumar, S., Ramaneswaran, S., Aneja, D., Jin, Z., Duraiswami, R. and Manocha, D. (2024) 'A Closer Look at the Limitations of Instruction Tuning', arXiv, available at https://doi.org/10.48550/arXiv.2402.05119 (accessed 11th March, 2024).
3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D. (2020) 'Retrieval-augmented Generation for Knowledge-Intensive NLP Tasks', in Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. and Lin, H. (eds), 'NIPS '20: 34th Conference on Neural Information Processing Systems', Curran Associates Inc., Red Hook, NY, pp. 9459–74.
4. Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S. and Chen, W. (2022) 'LoRA: Low-Rank Adaptation of Large Language Models', International Conference on Learning Representations.
5. Feng, W., Hao, C., Zhang, Y., Han, Y. and Wang, H. (2024) 'Mixture-of-LoRAs: An Efficient Multitask Tuning for Large Language Models', arXiv, available

at https://doi.org/10.48550/arXiv.2403.03432 (accessed 11th March, 2024).
6. Sheng, Y., Cao, S., Li, D., Hooper, C., Lee, N., Yang, S., Chou, C., Zhu, B., Zheng, L., Keutzer, K., Gonzalez, J. and Stoica, I. (2023) 'S-LoRA: Serving Thousands of Concurrent LoRA Adapters', arXiv, available at https://doi.org/10.48550/arXiv.2311.03285 (accessed 11th March, 2024).
7. Zhang, B., Change, D., Qian, E. and Agaby, M. (2023) 'Our Humble Attempt at "How Much Data Is Needed to Fine-tune"', available at https://barryzhang.substack.com/p/our-humble-attempt-at-fine-tuning (accessed 5th February, 2024).
8. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M. and Wang, H. (2023) 'Retrieval-augmented Generation for Large Language Models: A Survey', arXiv, available at https://doi.org/10.48550/arXiv.2312.10997 (accessed 11th March, 2024).
9. Ma, X., Gong, Y., He, P., Zhao, H. and Duan, N. (2023) 'Query Rewriting for Retrieval-augmented Large Language Models', arXiv, available at https://doi.org/10.48550/arXiv.2305.14283 (accessed 11th March, 2024).
10. Abdullah, N., Feizollah, A., Sulaiman, A. and Anuar, N. (2018) 'Challenges and Recommended Solutions in Multi-Source and Multi-Domain Sentiment Analysis', *IEEE Access*, Vol. 7, pp. 144957–71.
11. Baruni, J. S. and Sathiaseelan, J. G. R. (2020) 'Keyphrase Extraction from Document Using RAKE and TextRank Algorithms', *IJCSMC*, Vol. 9, No. 9, pp. 83–93.