

Harnessing AI for Business Insight: Key Considerations for Deploying LLMs in Summarization Pipelines

Alex P. Miller, Kartik Hosanagar, Ramayya Krishnan

The Oxford Handbook of the Foundations and Regulation of Generative AI

Philipp Hacker (ed.) et al.

<https://doi.org/10.1093/oxfordhb/9780198940272.013.0045>

Published: 21 August 2025

Abstract

Large language models (LLMs) have demonstrated significant potential in handling unstructured, natural language data. However, their adaptation to complex business settings remains a challenging endeavor. We explore the challenges and research opportunities associated with deploying LLMs for document and knowledge summarization in various business applications. We evaluate current paradigms for evaluation and highlight their inability to fully capture the multi-dimensional considerations in business settings, including relevance, provenance, and factuality of the output. The chapter emphasizes the need for a paradigm shift in evaluation approaches to better align with the nuanced needs of business applications. Key considerations include expanding the dimensionality of automated evaluation metrics, incorporating human-computer interaction factors, and addressing domain-specific needs. A case study on summarizing user-generated content from a product announcement video on social media is presented to illustrate these challenges and associated research opportunities.

Keywords: [large language models](#), [LLM](#), [LLM evaluation](#), [LLM summarization for decision support](#), [social media analysis](#), [user-generated content \(UGC\)](#), [artificial intelligence](#), [AI](#), [AI-driven business intelligence](#)

Subject: [Company and Commercial Law](#), [Law](#)

Series: [Oxford Handbooks](#)

Collection: [Oxford Handbooks Online](#)

I. Introduction

Many developments in machine learning and AI in recent decades have found meaningful business applications. Document retrieval and search techniques, content and product recommender systems, and supervised learning algorithms that use labeled data to predict human behavior are widely used in business-critical contexts and have all proven to be useful across many sectors of the economy.

Recent advancements in deep learning, particularly the advent of the ‘generative pretrained transformer’ and associated technologies, have generated a new wave of interest in AI [1]. Because of their ability to accept and generate unstructured, natural language, and multi-modal input from humans, new generative AI technologies have drawn considerable attention from investors, managers, media, and consumers. The impact of these changes on business and society is likely to be very significant. It has been argued that machine learning technologies in general, and generative pretrained transformers in particular, have the makings of general-purpose technologies [2, 3].

While generative AI has opened up new paradigms for working with text-based documents and data across a wide range of businesses and organizations, it remains unclear precisely if and how language models can be adapted to meet organizational needs, especially in high-stakes contexts. Commentators have speculated about the significant potential of generative AI technologies across many sectors of the economy, yet the path between existing versions of these tools and economically viable business models and applications is rarely explicitly described.

The most accessible and widely marketed generative AI tools (e.g. OpenAI’s ChatGPT, Microsoft’s Copilot applications, or Google’s Gemini chatbot) excel at generating textual outputs that faithfully respond to input prompts and commands. At the same time, each of these tools displays messages about the outputs of their models, in which their publishers are sure to disclaim the outputs that they generate (see Table 1). However, most businesses cannot simply disclaim the outputs of their LLM-enabled production processes, and improving their accuracy and fidelity continues to be an important challenge for their widespread use by organizations.

Table 1. Disclaimers Shown to Users in Popular Generative AI Interfaces

OpenAI	‘ChatGPT can make mistakes. Consider checking important information.’
Microsoft	‘Copilot uses AI. Check for mistakes.’
Google	‘Gemini may display inaccurate info, including about people, so double-check its responses.’

While computer scientists and AI engineers continue to invest in the general-purpose capabilities of next-generation foundation models, we believe there is a significant amount of value to be found in researching ways to leverage the capabilities of existing AI models for task and domain-specific applications while addressing limitations of the underlying foundational models. For text-based tasks whose outputs are intended to be read and acted upon by human agents, the designers of tools must engage with the real-world tradeoffs as well as user experience and human-computer interaction elements of the task.

We examine these considerations in the context of summarization applications built on top of foundational models. Summarization of text is one of the most popular applications of LLMs, with applications in a range of settings including summarization of news stories, books, legal contracts, social media comments, and more. We will refer to summarization in the broadest sense of the term for now, bearing in mind the importance of domain-specific adaptation. There do exist commonalities across the various applications of summarization technologies that are worth discussing at a high level before exploring context-specific considerations.

We will first discuss existing paradigms for evaluating summarization algorithms. We then highlight how, while the progress made based on these paradigms has been surprising and impressive, there remain important shortcomings in terms of their use in business contexts. We then focus the bulk of this article on describing how these challenges are being handled in emerging research and business applications. Next, we will ground our discussion in a real-world context with a case study that explores key considerations for applying LLMs to the task of summarizing user-generated content (UGC) on large social media platforms. We will conclude with a discussion about future research directions for AI and business researchers to explore.

II. Evaluation of Summarization Algorithms

A. Existing Paradigms

We first describe existing paradigms for evaluating summarization algorithms in computer science. One of the most widely used metrics for assessing the quantitative performance of summarization algorithms over the past 20 years is ‘Recall-Oriented Understudy for Gisting Evaluation’ or ‘ROUGE’ [4]. Prior to the development of ROUGE, the evaluation of machine-generated summaries was largely subjective, relying on ad-hoc human judgment. The simplest and most straightforward way to judge the quality of a summary is to manually compare a summarized text to the original text which it summarizes, using human judgment to determine whether a summary captures the main and most important points. However, doing this comparison by hand for computer-generated summaries is extremely time-consuming, especially with large amounts of text.

The advent of the ROUGE metric represented a significant development in automated textual analysis. The logic behind the metric is simple enough to explain: Assuming one has a ‘gold-standard’ human-written summary for a given body of text, ROUGE provides a simple numerical measure of the overlap in content between a machine-generated summary and the reference summary. More specifically, this measure of ‘overlap’ is based on the average (harmonic mean) of two sub-measures of textual overlap derived from the statistical notions of ‘recall’ and ‘precision’. In this context, recall measures how much of the relevant information in the reference summaries is captured by the machine-generated summary, while precision measures how much of the information in the machine-generated summary is also in the reference summary.¹

One limitation of ROUGE-based metrics is that they rely solely on exact matches of words, phrases, or n -grams to measure information overlap. It is possible for two summaries to be semantically very similar without a high degree of overlap in the actual words or phrases in the summaries. Additionally, for creative tasks like story or news summarization, two summaries can have high overlap in their n -grams but may share very different interpretations of the original text. To address this, researchers have turned to contextual embeddings (i.e. numerical representations of text designed to capture linguistic context and meaning in ways that are less sensitive to exact phrasings) [5]. This approach has led to the adaptation of general text evaluation metrics, such as BERTScore [6] and MoverScore [7], for the purposes of evaluating machine-generated summaries in ways that are more flexible than n -gram-based approaches [8, 9].²

B. Considerations for Broad-based Summary Evaluation

While the aforementioned metrics have been essential for the advancement of NLP research, as generative AI summarization tools make their way into more applied and high-stakes contexts, we contend the field will need a paradigm shift in how it thinks about evaluation. Ultimately, these metrics can only capture a small slice of the many considerations that are at play in summarization systems meant to meaningfully impact business practice. In these contexts, practitioners may naturally place high value on ensuring summarization algorithms consider factuality, density of information, and accurate reflections of textual tone and style. Yet when researchers critically probed the capabilities of existing summarization evaluation metrics, it is apparent these measures fail to capture many of these desired qualities. The shortcomings of ROUGE and BERTScore in particular are explored by Deutsch and Roth [14], who demonstrate that, rather than capturing meaningful information and syntactic meaning, these metrics are mostly limited to measuring the extent to which two pieces of text discuss the same topics. As a result, summarization algorithms that are exclusively optimized for these outcome measures will be limited by these shortcomings.

Particularly with the advent of the transformer neural network architecture and the recent success of large language models (LLMs), the need to think more carefully about the exact objectives of a summarization system is becoming more pressing. Researchers have already demonstrated that LLM-based algorithms are able to generate summaries of news articles that are judged to be on par with high-quality human-written summaries [15]. At the same time, existing automated metrics correlate poorly with human-rated judgments of quality for AI-generated news summaries [16]. This suggests the NLP research community is in need of more sophisticated and nuanced ways of assessing the quality of machine-generated summaries if the field is to continue advancing the capabilities of LLM-based summarization systems and bring them to high-stakes business use cases.

In thinking through the details of this nuance, we have identified three higher-order concerns that we believe will be crucial to the development of business-ready summarization systems. The first is increasing the dimensionality of automated evaluation metrics; the second set of concerns revolves around human-computer interaction factors; and the last element to consider is the importance of domain and context-specific factors. We discuss each of these in more detail and review related research. In Section III of this article, we will explore how these considerations manifest in the context of a specific business application of LLM-based summarization: summarizing UGC from a product announcement video. We will use this case study to illustrate how these considerations can be applied in a specific context that illuminates the broader challenges and opportunities in the field of LLM-based summarization.

1. Expanding Dimensions of Automated Quality Evaluation

Expanding the metrics used for automated summary evaluation by incorporating multi-dimensional real-world considerations is emerging as a consistent theme in recent research on applying language models for text summarization. In Kryściński et al. [17], a team of private industry researchers critiqued several dominant paradigms in automated text summarization and identified poor evaluation metrics as a key limiting factor in advancing the state-of-the-art. In line with our previous comments, the authors discussed how ROUGE metrics are limited by their reliance on exact-token matching and their inability to assess the factuality of generated summaries. In an attempt to move the field forward, the researchers propose four new dimensions of summarization quality: ‘relevance (selection of important content from the source), consistency (factual alignment between the summary and the source), fluency (quality of individual sentences), and coherence (collective quality of all sentences)’ (p. 6). A key takeaway from the work is that ROUGE metrics have only weak correlations with human ratings of text summaries along the aforementioned dimensions. In subsequent follow-up work, Fabbri et al. [18] studied how 14 different evaluation techniques (including several variants of ROUGE and other automated metrics) compared to expert-rated scores along the same four dimensions. They similarly conclude that existing automated metrics do not

reliably correlate with the multi-dimensional human evaluations. A related but different set of evaluation dimensions is discussed by Aharoni et al. [19], who proposed the following three dimensions of summary quality: informativeness (how well important information is captured by a summary), comprehensibility (ability to understand the summary), and attribution (factual alignment of summary with target article). They likewise found relatively weak correlation between ROUGE-based scores and their higher-order measures of summary quality. Taken collectively, these papers demonstrate that when specific dimensions of summary quality are of interest, the field is in need of more sophisticated evaluation techniques than those currently in use.

Even if some of the relevant dimensions are known, one of the major challenges is how to automatically generate evaluation scores for all the relevant dimensions, especially in the absence of a training dataset with human evaluations. Liu et al. [20] showed that it is possible to use LLMs with chain-of-thought prompting to generate reference-free metrics for evaluating text generated by LLMs. They showed that their LLM-generated metrics have reasonably high correlations with human evaluations and perform better than conventional metrics like ROUGE and BERTScore. However, they also point out a potential limitation that LLM-based metrics prefer LLM-generated text over human-written text.

If we are to accept that the objective of text summarization is inherently multi-dimensional, we must also give up the ‘myth of the single reliable number’. This phrase was coined by Gehrmann et al. [21], who argued that the lack of correlation among NLP evaluation metrics more generally suggests that no unidimensional objective function can fully characterize a natural language generation system. While expanding the number of dimensions in the evaluation of NLP summaries is a necessary step for moving the state-of-the-art forward, this also requires researchers to contend with the reality of inherent tradeoffs between dimensions, such as accuracy versus readability or completeness versus conciseness.

The need for more explicit evaluation of tradeoffs between evaluation dimensions is also acknowledged in the recent work of Adams et al. [22]. The authors proposed what they call a ‘chain of density’ prompting technique, which is designed to identify important entities in a target document and write an iterative series of summaries, with each successive summary containing more entities than the previous one while keeping summary length constant. In their experiments, the authors identified a tradeoff between the coherence/readability and informativeness of their summarization technique but left it for future work to more explicitly quantify and characterize the nature of this tradeoff. In total, we conclude that a key avenue of future work in this space is to continue developing more sophisticated evaluation techniques that can capture the multi-dimensional nature of summary quality.

2. Human-Computer Interaction Factors

As AI techniques have improved more generally in recent decades, automated tools have been inserted into a diverse set of organizational processes and systems. In many of these systems,

the results of automated tools are used as merely one input into a more complex process involving human judgments and interactions. This has led to the burgeoning of growth in research on the complex ways in which automated tools are used within social systems and organizations [23]. Almost by tautology, the output of any text summarization algorithm will be optimized for human consumption, presumably with the goal of increasing the efficiency, accuracy, or quality of a downstream task. Indeed, summarization algorithms are not designed to and have no capacity to make decisions or take any particular action. As such, in nearly any system that aims to make use of text summarization algorithms, human-computer interaction factors will also be important to consider when comprehensively evaluating the utility of such algorithms.

As the capabilities of NLP and text summarization techniques continue to advance, we believe many of the lessons from work on the interaction between humans and more classical machine learning models will be relevant for the field to consider. This will include insights from the social sciences about how humans think and reason, as well as how they interact with and interpret the outputs of AI systems [24, 25]. Research on the topics of ‘interpretability’, ‘explainability’, and ‘trust’, as they relate to human interaction with classification models, may also provide useful insights in generative AI settings. It is known that providing explanations about the internal logic of black-box classifiers is an effective technique for increasing trust in their outputs [26, 27]. However, some subtle and less obvious insights from this body of work may provide valuable lessons for NLP researchers. For example, Narayanan et al. [28] found that some forms of longer or more nuanced explanations diminish the experience and efficiency of experimental participants. Further, Bansal et al. [29] found across several tasks that giving users model explanations provides a null to minimal improvement in downstream decision accuracy beyond human-AI teams without explanations.³ Similarly, Ahn et al. [30] found that explanations had minimal impact on user trust and task performance for lay people using AI-based decision support.

Unsurprisingly, the intersection of HCI and NLP specifically is a growing area of research. Deutch et al. [31] developed a query system that is designed to explain how answers are derived from an existing structured database with detailed information about the provenance of the data used in its answers. The authors conducted a human-study and found that their system is able to increase assessments of the quality of the answers provided by the system, demonstrating in principle that explanations can be used to increase value in the outputs of AI systems. Further, a recent article by Schuff et al. [32] provides a detailed overview of conducting human evaluation studies in NLP research. Lastly, a 2021 tutorial on explainability in NLP provides a comprehensive discussion about the many issues and challenges in the field [33]. The authors specifically highlighted challenges around unifying the many different definitions of explainability, the need for context and subject-specific notions of explainability, and the importance of engendering trust in the users of AI systems.

The lesson from this body of work is that the outputs of AI systems are not merely a function of the quality of the underlying algorithms, which has been a dominant focus of the research community to date. Rather, human assessments of the outputs of AI systems are also a function of the ways in which humans interact with and interpret the outputs. This theme is also a key takeaway from early research on the use of LLMs in business settings, which found that while generative AI can improve human performance on a number of tasks, it is not always a complement to human labor [34]. This is largely due to the fact that humans end up trusting AI-generated text even on tasks where LLMs perform poorly. Helping human users of AI systems navigate the landscape of AI capabilities and effectively instilling *appropriate* levels of confidence and trust will be a key challenge as these systems become more widely adopted. As such, as we consider the future of using LLMs for summarization tasks, we suggest that researchers should carefully consider the broader literature on human-computer interaction and the social sciences when designing and evaluating summarization systems.

3. Context-Specific Requirements

When working with models of natural language, it is essential to consider the specific context in which the model will be used. Whereas classical machine learning models are narrowly focused on a single task in which the context is well-defined, LLMs are designed to be general-purpose models that can be applied to a wide range of tasks and domains. This means that the requirements of the task and the context in which the model will be used are likely to vary widely depending on the specific application.

This is particularly true for summarization tasks, which can vary widely in terms of the requirements of the task and the context in which the summary will be used. A summary intended for an academic audience, such as one summarizing a scientific paper, must focus on distilling complex findings and conclusions using technical language appropriate for experts. Conversely, a summary for a general newspaper reader should provide an accessible overview of the main events or arguments, free from jargon and technical details. This differentiation in content and style is vital, as the needs of the intended audience of a summary can vary uniquely from context to context.

Apart from the language or style of a summary, different domains may also have different desiderata for the informational content of a summary. For example, in some domains (e.g. law or medicine), it may be critical to minimize false negatives (i.e. missing important information), while in other domains (e.g. news or social media), it may be more important to minimize false positives (i.e. including irrelevant or false information). As such, the specific requirements of any given summarization task and the context in which the summary will be used are likely to have a significant impact on the design and evaluation of the model. This means that the considerations raised here about expanding the dimensions of quality evaluation and the human-computer interaction factors involved in a given task will also vary across contexts.

There are engineering techniques (e.g. fine-tuning and retrieval augmented generation) that have already been developed to facilitate context-specific adaptation of language models. We believe these methods will be important for leveraging LLMs in many domains. Yet early research shows that the specific implementations required to get the most out of these techniques is highly task-dependent, suggesting that context-specific factors and considerations will remain relevant to those who are applying these technologies for the foreseeable future [35].

To illustrate the importance of these considerations in a real-world business context, we will now turn to a case study of a specific application of LLM-based summarization. We believe this will make clear how the considerations discussed in this section materialize in a concrete business setting and provide lessons for future research and practice in the field of AI-generated summarization.

III. Case Study

Social media platforms have become an important channel for businesses to engage with customers and gather feedback on their products and services. Engaging with audiences on various platforms such as YouTube, Instagram, or Reddit is a common way for businesses to introduce new products to customers and generate excitement around their brand. Comments sections or replies to social media posts on brand-related content can potentially provide valuable insights into customer sentiment, preferences, and behavior. However, the sheer volume of comments on large social media platforms can make it difficult for businesses to manually read and analyze all of the comments. Summarization techniques, particularly with recent advances in the capabilities of language models, offer a potential way to distill the key insights from user-generated content (UGC), enabling businesses to make more informed decisions based on this data.

The downstream audience of a UGC summarization system may be anyone from a diverse group of stakeholders, including product managers, marketing teams, customer service representatives, and executives. At a high-level, the output of the summarization system should inform some set of decisions or actions that are downstream of the summarization process itself. This is illustrated in a high-level diagram depicted in Figure 1, which outlines the key steps involved in a generic summarization pipeline.

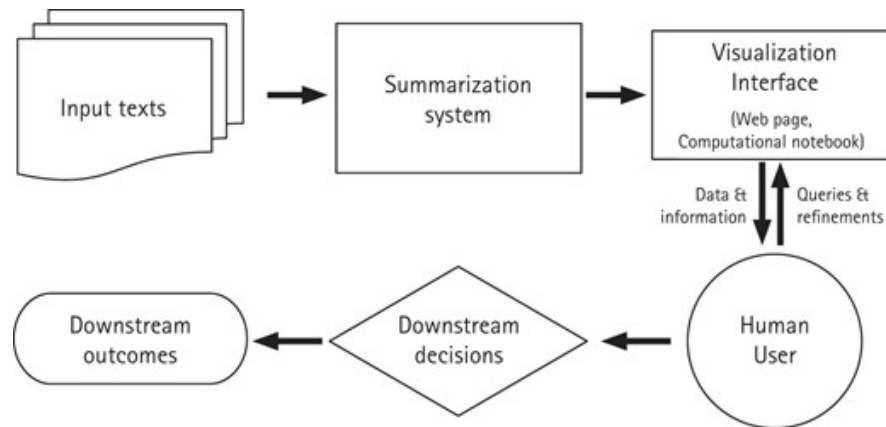


Figure 1. Outline of Generic Summarization Pipeline.

Notes: This diagram is meant to capture the key feature of a large class of summarization pipelines. Any summarization system that adds value in real-world contexts should necessarily inform some set of decisions or actions that are downstream of the summarization process itself.

For the purposes of this case study, we will focus on the task of summarizing UGC from a product announcement video, with the goal of extracting key insights from the comments section of the video for a product or marketing manager. As such, our system should be specifically designed to provide a concise summary of the main themes and sentiments expressed by users for the purposes of informing future product development and marketing strategies. This will give us the opportunity to explore the concrete implications of the considerations discussed in the previous section in a real-world context.

A. Background on Application Context

We will specifically be analyzing UGC in the form of YouTube comments that are reacting to the announcement of a recent smartphone (i.e. the ‘Phone 2a’) launched by Nothing Technology Limited.⁴ The video, uploaded to Nothing’s main YouTube channel on 5 March 2024, is a promotional commercial that describes the key features of the device and showcases its design and functionality. Typical of product announcement videos on YouTube, the commercial has garnered a number of comments from users expressing their reactions, opinions, feedback, and suggestions about the product. The comments section of the video provides a rich source of data that can be analyzed to understand the sentiment of users, identify key themes and topics of discussion, and extract actionable insights for product development and marketing strategies. Downstream business decisions that could be informed by these insights include product-feature enhancements, marketing campaign adjustments, customer service improvements, market-entry strategies, and competitive-positioning strategies.

Though this context is specific to a technology product announcement video on YouTube, the principles and techniques discussed in this case study also speak to a wider range of social

media platforms and user-generated content. Companies in many industries use social media platforms to announce and promote new products and engage with customers. We believe there is significant potential for these companies to leverage LLMs and summarization techniques to analyze UGC on social media and extract valuable insights for business decision-making. Even outside of the context of analyzing UGC from a product announcement video, the considerations we discuss in this case study will also help illustrate the broader challenges and opportunities in the field of AI-generated summarization.

Given our context, our goal is to develop a summarization system that can process the comments section of the video and generate a concise summary that captures the main themes and sentiments expressed by users. The summary should provide a high-level overview of the feedback, highlighting the key positive and negative aspects of the product as perceived by users. The system should be able to identify distinct themes and topics of discussion in the comments, group related comments together, and extract relevant details and insights from the feedback. The summary should be structured in a way that is easy to read and understand, providing a clear and concise representation of the main insights from the comments section.

Throughout this case study, we will be using OpenAI’s GPT-4 (accessed via its application programming interface) as the underlying language model for our summarization system. As of the writing of this article, GPT-4 is a state-of-the-art large language model that has been fine-tuned on a diverse range of text data and is capable of generating high-quality text across a variety of tasks and domains. We will use GPT-4 to process the comments section of the Phone 2a announcement video and generate a summary of the UGC. We will also explore different techniques and strategies for enhancing the summarization process and improving the quality and relevance of the summary output.

B. Baseline Approach

We begin by considering a relatively vanilla approach for using GPT-4 in our summarization pipeline. We will use this to provide a baseline against which we can describe the various ways it is possible to enhance the pipeline given the objectives of our context. In Table 2, we reproduce the system prompt that we have provided to the language model, which is designed to guide the system in identifying the main themes and sentiments expressed by users and generating a comprehensive summary of the feedback.

Table 2. Baseline Summarization Prompt

System	You are a text summarizing bot. Your primary purpose is to identify a comprehensive list of distinct high-level themes from text-based documents. For each theme, provide a name/title for theme and include a bullet list of details relevant to the theme drawn from the feedback.
---------------	--

Table 3 contains the output of the baseline summarization system, generated by GPT-4 in response to the system prompt provided in Table 2, with the {comments} placeholder in the user prompt being replaced with line-separated comments from the YouTube video.

Table 3. Output of Baseline Summarization

User	Please summarize the following texts: {comments}
Assistant response	<ul style="list-style-type: none">• Product Quality and Performance Issues<ul style="list-style-type: none">○ Concerns include manufacturing defects, unsatisfactory customer-care resolutions, and quality of build materials, notably plastic backs prone to scratching.○ Performance-related complaints focus on the need for bug fixes and updates for smoother operation, especially in gaming contexts.○ Mixed reviews on processor choice, comparing MediaTek unfavorably with Snapdragon, alongside requests for future models to have better water resistance and display options.• Design, Aesthetics, and Build Quality<ul style="list-style-type: none">○ General appreciation for the phone’s aesthetic, with specific praise for distinctive features like the Glyph Interface. However, there’s feedback indicating a desire for a more compact design, color options, and premium materials such as frosted glass.○ Concerns regarding the susceptibility of the design elements (like the camera area) to scratches.• Software Features and User Interface<ul style="list-style-type: none">○ Requests for updates focusing on camera performance improvement and the inclusion of missing features.○ Suggestions for additional software tools and improvements to the user interface for better user experience.• Pricing, Market Feedback, and Accessibility<ul style="list-style-type: none">○ Discussions around the phone’s price point, with some suggesting it’s beyond budget or overpriced, and requests for price reductions.○ Comparative analysis with competing brands, highlighting demands for the phone to offer better specs or pricing to remain competitive.○ Calls for broader market availability and better after-sales support, including specific requests for the phone to be available in more regions and countries.• Customer Support and Brand Perception<ul style="list-style-type: none">○ Frustration expressed with warranty terms and the need for better after-sales support.○ Overall, the brand’s unique market position and distinct design philosophy have been appreciated, with excitement for future products. However, there’s skepticism regarding the brand’s long-term viability and criticisms of its marketing strategies.

Overall, we can see that the baseline approach provides a useful summary of key themes and sentiments expressed in the comments section. Without any particular guidance, the system has successfully identified distinct themes such as product quality and performance issues, design, aesthetics, and build quality, software features and user interface, pricing, market feedback, and accessibility, and customer support and brand perception. For each theme, the system has provided a bullet list of details relevant to the theme drawn from the feedback. This summary provides a comprehensive overview of the main insights from the comments section and highlights the key aspects of the user feedback that are relevant for product development and marketing strategies.

While we should not forget how impressive it is that any fully automated system can generate such a summary, several factors about this technique limit its utility in business contexts where the goal of the summarization pipeline is to use its insights to inform consequential business decisions. We will now review some of these limitations and use the factors discussed in the previous section to help illustrate where the current state-of-the-art stands and where future research efforts should be focused.

C. Moving Beyond Baseline

1. Additional Metrics for Evaluation

The extra dimensions of quality that we identified in the general literature, such as relevance, consistency, fluency, and coherence, are relevant in our context. Automated metrics such ROUGE require a gold-standard dataset of human-generated reports that can serve as the reference text against which LLM-generated text can be compared. Generating such a gold-standard dataset can be expensive. Furthermore, ROUGE and related metrics have been shown to have a low correlation with scores generated by human evaluators.

One potential solution proposed [20] is to use LLMs as evaluators of natural language text generated by an LLM. In particular, if we can identify some of the relevant dimensions along which to evaluate the summary report generated by an LLM, we can construct prompts for an LLM to score the output of our system. Inspired by Liu et al. [20], we have provided sample prompts in Table 4 on multiple dimensions; these can be used to develop a more comprehensive set of evaluation metrics for the system in our context.

Table 4. Prompts for Multi-Dimensional Evaluation of the Output

Coherence	<p>You will be given a managerially actionable summary of social media comments posted in response to a product announcement. Your task is to rate the summary on one metric.</p> <p>Evaluation Criteria</p> <p>Coherence (1-5)—the collective quality of all sentences. The summary should be well-structured and well-organized. The summary should not just be a heap of related information but should build from sentence to a coherent body of information about a topic.</p> <p>Evaluation Steps</p> <p>1. Read the social media comments carefully and identify the main topic and key points. 2. Read the summary report and compare it to the comments. Check if the summary covers the main topic and key points, and if it presents them in a clear and logical order. 3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria. Source Text: {comments} Summary:</p>
Consistency	<p>You will be given a managerially actionable summary of social media comments posted in response to a product announcement. Your task is to rate the summary on one metric.</p> <p>Evaluation Criteria</p> <p>Consistency (1-5)—the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. This criterion penalizes summaries that contain hallucinated facts.</p> <p>Evaluation Steps</p> <p>1. Read the social media comments carefully and identify the main facts and details it presents. 2. Read the summary report and compare it to the article. Check if the summary contains any factual errors that are not supported by the article. 3. Assign a score for consistency on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria. Source Text: {comments} Summary:</p>

2. Nuances of Quality

Given that the purpose of the analysis is to inform particular aspects of product development and marketing strategies, the ability of the system to accurately identify *actionable* insights from the comments is critical for a marketing manager to make informed decisions. In particular, managers would likely be keenly interested in understanding the overall valence of audience reactions.

Notice how the baseline approach includes both positive and negative feedback under the ‘Design, Aesthetics, and Build Quality’ topic. This is a reasonable approach for a general-purpose summarization system, but in a business context, one may be concerned that some negative feedback is lost in the noise of positive feedback. This is particularly important in the context of a product announcement video, where the product manager is likely to be interested in identifying areas for improvement or potential issues that need to be addressed. A system in this context should ideally be able to accurately capture the sentiment of the comments and provide a summary that reflects the overall sentiment of the audience. This is a key dimension of the value of the output that is specific to the manner in which the summary is used in downstream tasks in this business context.

One way to address this issue is to explicitly incorporate sentiment analysis into the summarization pipeline. In particular, if we know ahead of time the broad categories of feedback (positive or negative in this case), we can use this information to guide the summarization process. In the context of our case study, we can provide additional system prompt text to guide the system in generating summaries that focus specifically on the positive or negative aspects of customer feedback. This will help ensure that the summary captures the overall sentiment of the audience and provides a more actionable overview of the feedback. We have provided a sample prompt (see Table 5) that can be used to guide the development of a system that incorporates sentiment analysis into the summarization pipeline.

Table 5. Additional System Prompt Text for Positive/Negative Summarizers

Positive Summarizer	Negative Summarizer
Focus specifically on the positive aspects of customer feedback. What did users like? What features do they appreciate? What aspects of the product stand out and generate positive sentiments? If there are some comments that express enthusiasm or a positive sentiment without much specific content about the product itself, include ‘General excitement’ or a similar theme in your summary. Omit any feedback suggesting dissatisfaction or negative sentiment. Omit any suggested changes or comments on how to improve the product.	Focus specifically on the negative aspects of customer feedback. What did users dislike? What features do they wish were different? What aspects of the product stand out for generating negative sentiments? Ignore any comments expressing excitement or pleasure; omit any topics pertaining to positive feedback. Focus on suggested changes or comments on how to improve the product.

We have provided the output of the sentiment-specific summarizers, generated by GPT-4 with the system prompt modifications. As can be seen in Table 6, the positive summarizer focuses

on the positive aspects of customer feedback, highlighting the features that users appreciate and the aspects of the product that generate positive sentiments. The negative summarizer (see Table 7), on the other hand, focuses on the negative aspects of customer feedback, highlighting the features that users dislike and the aspects of the product that generate negative sentiments. Note how different aspects of design appear in both the positive and negative results. By separating out the positive and negative feedback, the system is able to provide a more nuanced treatment of both aspects of user feedback. In theory, this should provide a more actionable overview for downstream users of the system, such as product managers or marketing teams, who are interested in understanding the overall sentiment of the audience and identifying areas for improvement or potential issues that need to be addressed.

Table 6. Themes from Summarizer Focused on Positive Feedback

<ul style="list-style-type: none">• General Excitement and Brand Affection: High levels of excitement and anticipation for the new “Nothing 2a” smartphone. Loyalty and pride in the Nothing brand’s journey, vision, and marketing approach are prevalent among comments. A community-like affinity towards the brand, highlighting a strong emotional connection.• Design, Aesthetics, and Unique Selling Propositions: The “Nothing 2a” is acclaimed for its unique and innovative design. Specific praise for the Glyph Interface and the overall see-through design. Recognition of unique features such as the Glyph Interface and overall design philosophy as major factors attracting consumer interest.• Performance and Software: Positive feedback on the device’s performance, especially its smooth operation and battery life. Appreciation for the Nothing OS for its user experience.• Pricing and Value: The phone’s competitive pricing is widely praised for offering excellent value.• Market and Product Availability: Significant anticipation for the “Nothing 2a” smartphone’s release in specific global markets. Excitement specific to the market launch in India, indicating the importance of geographic expansion strategies.• Product Features and Specifications: Interest in high-capacity battery, fast charging capabilities, and “Volume increasing light”.• Camera Quality: Some praise for the camera system’s capabilities.

Table 7. Themes from Summarizer Focused on Negative Feedback

-
- Privacy and Data Security: concerns over user data privacy and trust issues regarding company handling.
 - Customer Service and Support: poor after-sales service, with reports of difficulties in issue resolution and obtaining refunds for DOA units.
 - Durability and Design: Absence of water and dust resistance criticized for lack of durability. Suggestions for design improvement include a preference for a dot matrix small display and use of more durable materials than plastic to prevent scratches. Specific dislikes regarding size, with calls for more compact options, as well as material quality concerns.
 - Performance and Hardware Specifications: Disappointment with the choice of MediaTek processor over a Snapdragon variant, highlighting performance concerns. Software bugs and FPS drops in games signal a need for software optimization. Critiques on using UFS 2.2 storage and LPDDR4X RAM as outdated, with a preference for UFS 3.1 or higher and LPDDR5 RAM for improved performance.
 - Camera Quality and Features: General dissatisfaction with camera performance, including inconsistency and lack of optimization. Requests for a cinematic video mode, better portrait mode, and protection for the camera area from scratches. Specific suggestions for increasing sensor size, retaining the camera layout, and a preference for a portrait lens over an ultra-wide lens.
 - Software and User Interface: negative feedback on Nothing OS for lacking features and unreliable performance, alongside a desire for more years of Android updates.
 - Pricing and Value: discontent with the device price, perceived as too high, especially concerning the additional cost of charge adapters not included in the package.
 - Additional Features and Preferences: Requests for headlining features like Dolby Atmos, Dolby Vision, a better fingerprint sensor placement, eSIM support, and the inclusion of a charger in the box. A notable preference for headphone jack availability and ergonomic device options. Feedback on packaging not being collectible or reusable, impacting the unboxing experience.
 - General Disappointment: expressions of overall disappointment focusing on various aspects like design choices, performance issues, and camera quality without detailed improvement suggestions.
-

3. Human-Computer Interaction Factors

Outside of the quality or characteristics of summarization algorithm itself, the way in which the summary is presented to the end user is also a critical factor in the utility of the summarization system.

One way to improve upon the baseline chat or API-based interfaces of existing LLMs would be to provide a user-friendly visualization of the language model's analysis of the comments. This could include features such as the ability to filter the summary by sentiment or theme, quantify the amount of feedback relating to each topic in the summary, and view detailed comments associated with each theme. By providing a structured and interactive visualization

of the summary, the system can help the end user quickly identify key insights and trends in the feedback and make informed decisions based on this data.

One area of concern that arises from LLM hallucinations is whether the LLM-suggested actions are derived from actual user comments and whether those comments are representative. Any information that allows a user to better understand the underlying comments behind the main themes identified by the LLM can be particularly useful. One step in this direction would be to develop a mapping between the themes identified by the summarization system and the actual content of the comments themselves. This would allow the marketing manager to see the actual comments that are driving the summary and provide additional context and nuance to the insights provided by the system.

To this end, we have provided a sample prompt (see Table 8) that can be used to guide the development of a system that incorporates sentiment analysis into the summarization pipeline. This can be used in tandem with an LLM’s formatting options to generate structured data for downstream visualizations and user-interface systems.

Applying this labeling system to the raw UGC enables a much richer set of downstream presentation modalities than the output of the LLM itself. For example, with the labels mapped to the comments using the prompt in Table 8, it becomes possible to provide users with a specific set of comments, rather than mere summaries. In Table 9, we have shown an example of how the comments can be matched to the summarization topics.

Table 8. Prompts for Topic-Labeling System

System prompt	<p>You are a text-labeling algorithm. You will be provided with a list of topics and a comment. For each topic, your task is to determine if the comment is related to it. A comment can be related to multiple topics.</p> <p>[+ formatting instructions]</p>
User prompt	<p>Considering the following topics</p> <p>TOPICS: {topics}</p> <p>Determine if the comment below is about or related to the preceding topics.</p> <p>COMMENT: {comment}</p>

This structured data can then be used in downstream visualizations and user-interface systems to provide a more interactive and informative experience for the end user. In the system we are discussing here, the structured data can be used to generate visualizations that show the distribution of comments across different topics, the overall sentiment of comments

related to each topic, and the specific comments that are related to each topic. Each of these visualizations can provide valuable insights into both the overall feedback from users and help the marketing manager make their own assessments of the success of the summarization algorithm in capturing the key themes and sentiments expressed in the comments section. This last point is crucial for the successful deployment of LLMs in business contexts. While the output of LLMs can be valuable and insightful, it is also unpredictable and known to make mistakes. By providing users with the ability to see the raw data that the LLM is working with, they can make their own assessments of the quality of the output and identify any potential errors or biases in the system. This can help build trust in the system and ensure that the end user is able to make decisions based on a realistic understanding of the capabilities and limitations of the LLM output.

As for how all the pieces of the system fit together, we have provided a diagram in Figure 2 that outlines the structured UGC labeling pipeline we have described in this section. This diagram shows how the raw text content is passed through our summarization pipelines to extract themes of a particular type (positive and negative, in this case), which are then passed through a labeling prompt to identify specific comments related to each theme.

Table 9. Sample of Comments Matched to Summarization Topics

Topic Name	Comments Related to Topic
Positive Topic: Design, aesthetics, and unique selling propositions	<ul style="list-style-type: none">• Yo, this phone looks sick!• This one’s actually beautiful.• I love this design. Nothing is the only company that makes phones that have character, and this is no exception.• Love or hate the product, the video is clean• This reminds me of walkman• I wish it was a bit smaller
Negative Topic: Performance and hardware specifications	<ul style="list-style-type: none">• Give me this phone with Qualcomm snapdragon 8 gen series• why not Snapdragon 8 gen 3 or Mediatek dimensity 9300 Ultra• They need to release a 512GB version of the Nothing 2A for around \$600 for Aussie fans since they are offering 3 OS updates• UFS 2.2 is not commendable Nothing, why not 3.1 why why??• is this an upgrade or downgrade to the phone 2?• How good is it in gaming does it deliver constant fps

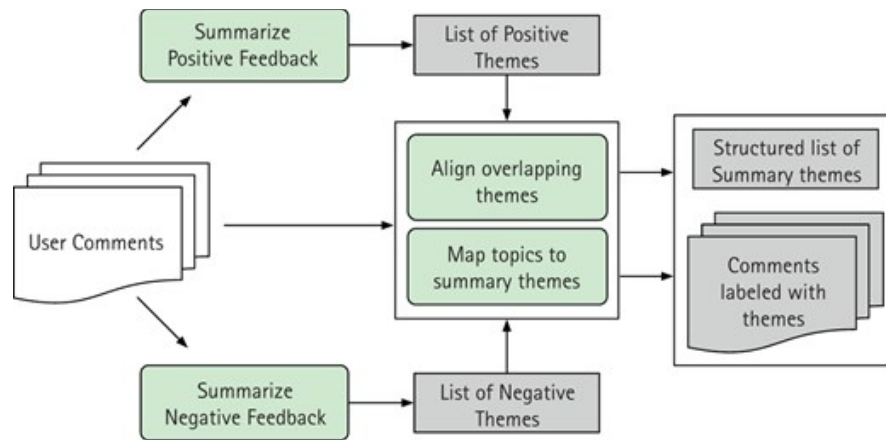


Figure 2. Outline of Structured UGC Labeling Pipeline.

Notes: The system laid out here describes an example of how one might send raw text content into multiple summarizers pipelines (left side). Once a set of themes is extracted from each summarizer, we can then pass these themes into a set of prompts that are designed to (a) identify when themes from each pipeline are about the same topic and (b) identify which themes apply to which raw text inputs (middle pane). The output of this process is structured data that can then be used in downstream visualizations and user-interface systems.

A key step in this system we have yet to discuss is the aligning of overlapping themes from the two summarizers. In many contexts, the themes extracted from different summarizers may overlap or be related in some way. For example, in our case, note how the themes of ‘Design, Aesthetics, and Unique Selling Propositions’ from the positive summarizer and ‘Durability and Design’ from the negative summarizer are quite similar. Even though they are framed differently, they are both related to the same product dimensions. In such cases, it is useful to have a mechanism for aligning these themes and identifying when they are about the same topic. This can help users get a sense of both the quantity of feedback related to a particular topic and the overall sentiment of that feedback. We have provided an example prompt for aligning topics (see Table 10) that can be used to guide the system in identifying when themes from different summarizers are about the same topic.

Table 10. Topic-Aligning Prompt

System	Determine if the following topics are nearly identical, very closely related, somewhat related, polar opposites, or completely unrelated to the context. Two topics should be considered
prompt	‘NEARLY IDENTICAL’/‘VERY RELATED’ if they are about the exact same subject or concept. If the topics are only somewhat related or have a substantial amount of overlap, they should be labeled ‘SOMEWHAT RELATED’. If the topics are complete opposites—as in, one topic is about the positive aspects of some brand/product/feature, and the other is about the negative aspects—they should be labeled ‘OPPOSITES’. If the topics are completely unrelated, they should be labeled ‘UNRELATED’.

Notes: This prompt is designed to help the system identify when two topics are about the same subject or concept. If two topics from different summarizers are about are deemed to be related in some way, our system then makes one more call to the language model to come up with a common label for the two topics. For example, applying this prompt to the topics ‘Design, Aesthetics, and Unique Selling Propositions’ and ‘Durability and Design’ were deemed to be ‘somewhat related’; our system then generated a new label ‘Design, Aesthetics, and Durability’ to represent both topics. Similarly, the positive topic of ‘General Excitement and Brand Affection’ and the negative topic of ‘General Disappointment’ were deemed to be ‘opposites’; the system generated a new label ‘Brand Sentiment’ to represent both topics.

The output of the summarizing system described in Table 10 consists of two primary components: the structured data that maps comments to themes and the structured data that maps related themes to each other. This structured data can then be used in downstream visualizations and user-interface systems to provide a more interactive and informative experience for the end user. We have provided an example visualization based on the results of the sample system we developed to analyze the UGC from the Phone 2a announcement video. This visualization is shown in Figure 3, which provides a color-coded representation of the average sentiment about each topic and a bar chart showing the number of comments related to each topic. This visualization provides a clear and concise overview of the main themes and sentiments expressed in the comments section and helps the marketing manager quickly identify key insights and trends in the feedback.

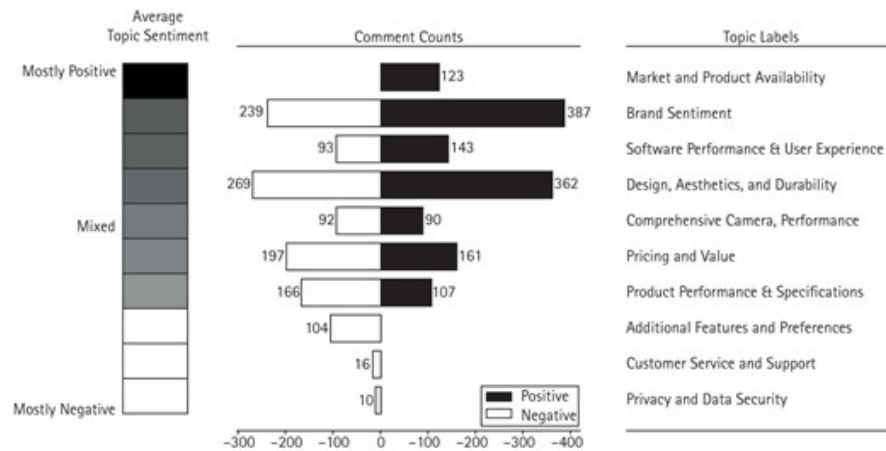


Figure 3. Example Visualization.

Notes: This visualization, derived from the outputs of the system described in Figure 2, is an example of how one can provide quantitative insight into the qualitative output of a language model summary. In particular, the leftmost pane features a color scale that represents the average sentiment about a particular topic, while the bar charts in the middle give the user a sense of how many individual comments are related to each topic. In a fully realized system, the structured data generated by the summarization system could be used to generate a more interactive user experience. For example, users could filter the summary by sentiment or theme, view detailed comments associated with each theme, and explore the relationships between different themes. This would provide a more engaging and informative experience for the end user and help them quickly identify key insights and trends in the feedback, while also providing a more nuanced understanding of the data and the limitations of the summarization system.

4. Business Context and Domain-Specific Considerations

Finally, let us discuss context-specific considerations that are relevant for getting the most out of the LLM-generated summaries in this setting. The first factor we highlight is how the prompts within a particular summarization pipeline can likely benefit from additional information about the nature of the pipeline itself. As one example, consider how each step in a summarization pipeline that relies on LLMs can likely benefit from being provided context about the pipeline itself. In the system we have been discussing, we have two separate summarizers, one focused on positive sentiment and the other on negative sentiment. We also have a step where we use a language model as a zero-shot classifier for assigning individual comments to the topics generated by these summarizers. However, the baseline labeling prompt described in Table 8 can be quite permissive in determining which topics it considers to be ‘related’ to a given comment. Note how in Table 8, the comment ‘I wish it was a bit smaller’ was derived from the summarizer focused on positive sentiment; also the comment ‘How good is it in gaming does it deliver constant fps’ was labeled as negative, despite being

more of an inquiry rather than a particular criticism about these features. Adding information to the labeling prompt that provides more context about the summarization pipeline itself (e.g. whether the labeler should only consider comments of a particular type (positive/negative)) can help the system make more accurate decisions about which comments are related to which topics (see Table 11 for an example). Providing this additional context can help ensure that each step in the summarization pipeline is working together effectively to generate a high-quality summary of the UGC.

Table 11. Additional System Prompt Context for Comment Labeling

System	A topic should only be deemed <i>related</i> if the comment has a positive sentiment toward the topic. If the comment expresses a negative sentiment or critical feedback, it should NOT be
prompt	labeled as related to the topic. Do not label any comment that is solely expressing a desire for a different feature or a change to the product in question.

In addition to this ‘inside view’ of the summarization pipeline, there are also ways to provide the LLM with additional context about the business context in which the summarization system is being used. For example, in our case, the summarization system is being used to analyze UGC from a product announcement video. In this context, it may be useful to provide the LLM with information about the source of the target texts and the typical characteristics of the feedback that is likely to be found in these texts. YouTube comments are known to be of highly variable quality and relevance; explicitly telling the LLM where the comments come from and ignoring common forms of spam or noise can be a useful way to improve the quality of the output (e.g. comments that are just emojis or ‘first!’). Further, the LLM can be told to ignore comments that are replies to other comments, as these may be less relevant to the main themes of the video. (In other contexts, such replies may be very relevant and should be included in the analysis.)

Other types of contexts that could be useful to provide to the LLM include information about the product itself, such as its specifications, features, and unique selling points. This information can help the LLM generate summaries that are more accurate and informative by providing additional context about the product and its key attributes. Further, if our summarization pipeline were intended to be used to analyze smartphone or technology product announcement videos more generally, it would be useful to describe the typical dimensions of evaluation in this industry category. In such a situation, we could steer the LLM toward common evaluative criteria such as internal hardware specifications, software interface, external materials and aesthetic design, and camera quality. While essentially all of these topics emerged from our initial analysis, the baseline approach faced some struggles in delineating between specific dimensions of feedback. For example, the topics of ‘Product

Quality and Performance Issues’ and ‘Design, Aesthetics, and Build Quality’ feature several related points of overlapping feedback. In our experiments, providing these additional context cues significantly helped the LLM generate more organized summarization themes with less overlap between topics.⁵

Finally, it is important to consider the broader business context in which the summarization system is being used. In our case, the system is being used to analyze UGC from a product announcement video, with the goal of informing product development and marketing strategies. In this context, it is important to provide the LLM with information about the business objectives of the analysis and the decisions that the summary will inform. This can help the LLM generate summaries that are tailored to the specific needs of the business and provide insights that are actionable and relevant to the end user.

The challenge of providing the LLM with this additional context is that it requires a deep understanding of the business context in which the summarization system is being used and the specific needs of the end user. This can be a complex and time-consuming process that may require input from domain experts and stakeholders. Naturally, this limits the scalability of any single summarization pipeline and may make it difficult to adapt the system to new business contexts or use cases. In some situations, it may be possible to automatically fill in additional context factors from the web or other sources. For example, the brand of the product could be automatically extracted from the video metadata, or the product category could be determined from the video title or description.

For more general-purpose summarization pipelines, it may be worth the effort to develop an ontology of common contexts and use cases. This ontology could be leveraged to guide the development of summarization pipelines in new domains by providing a set of common context types and prompts. This could help strike a balance between the flexibility and power of LLMs and the nuance and context needed to leverage them in task-specific settings. In Table 12, we have provided a basic outline of the high-level types of contexts that could be useful as a starting place for developing an ontology for use in summarization pipelines revolving around product feedback more generally.

Table 12. Relevant Dimensions of Outside Context in Summarization Pipelines

Context Type		Description
Source	and characteristics of target text	Information about the source of the text, such as the platform or website where the text was found, and common types of text to highlight versus ignore.
Product information		Information about the product itself, such as its specifications, features, and unique selling points.

Product details	category	Information about the typical dimensions of evaluation in the product category, such as external materials and aesthetic design, internal hardware, software interface, and camera quality.
Brand information		Information about the brand or company, such as its history, values, and existing product lineup.
Business context		Information about the business context in which the summarization system is being used, such as the objectives of the analysis and the decisions that the summary will inform.

One could also consider expanding the set of context topics to include recent news articles, information about relevant competitors, or even the current state of the market. These additional categories add a significant amount of complexity to the system, but—depending on the exact problem at hand—they could be crucial for generating high-quality summaries that are actionable and relevant to the end user. Overall, however, it is important to strike a balance between providing the LLM with enough context to generate accurate and informative summaries and overwhelming the system with too much information. This is a key challenge that must be addressed in the development of summarization pipelines for business contexts and will likely require input from domain experts as well as experimentation and iteration to find the right tradeoffs.

IV. Conclusion and Future Directions

We have only covered a subset of the considerations that are relevant to improving the quality of LLM-generated summaries in general business contexts. However, this case study has illustrated what we believe are some of the key ways in which the output of LLM-generated summaries can be better leveraged in real-world applications. Many aspects of the challenges we discussed here can be considered representative of a large class of summarization pipelines that are intended to inform consequential decisions. We specifically highlighted how considering the specific dimensions of quality that are relevant to the business context, thinking carefully about the presentation of the summary to the end user (both in terms of raw output but also in terms of quantification, attribution, and visualization), and providing the LLM with additional context about the summarization pipeline, the business context, and the domain-specific considerations can all help to ensure that the LLM-generated summaries are accurate, informative, and actionable.

The case study we have presented here is merely meant to point to the potential for LLMs to be used in real-world business contexts and the challenges that must be addressed to make this a reality. The specific techniques and strategies we have discussed are just a starting point, and there is much more work to be done in this area. Future research efforts may be most valuable

by focusing on the following lines of inquiry: (1) Conducting more in-depth experimentation within a specific real-world context, including studies of how end users respond to different design choices for the summarization system. (2) Fine-grained evaluation based on subsets of data. This is often needed to identify social biases or other systematic errors in the output of AI systems (e.g. how an LLM-based career counseling app recommends career options for male versus female students). Cabrera et al. [36] presented a framework for behavioral evaluation of ML systems that combines a Python API with an interactive UI to visualize metrics and write unit tests. Additional work on fine-grained behavioral evaluation, as well as on better visualization of evaluation results, will be valuable. (3) Developing more sophisticated techniques for improving the quality of LLM-generated summaries in business contexts, such as incorporating purpose-built sentiment analysis or topic modeling algorithms. (4) Developing more interactive and informative user interfaces for working with LLM summaries and determining which elements of the user interface can be generalized across business contexts. (5) Studying the ontology of common summarization paradigms and delineating different types of pipeline architectures that are most effective in different business contexts. This last point is particularly important for the future of LLMs in business contexts, as it will be crucial to find the right balance between the general models' power and flexibility and the nuance and context needed to leverage them in task-specific settings. The power of LLMs is immense and particularly well-suited to the challenges of summarization, but the path to realizing this potential in business contexts will require developing products and interfaces that go beyond mere chat or API-based systems if they are to be used widely by the greater business and professional communities. We hope that the case study and discussion presented here will serve as a starting point for future research efforts in this area and help to guide the development of more effective and useful LLM-based summarization systems for business applications.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 30 *Advances in Neural Information Processing Systems* 6000–6010. <https://dl.acm.org/doi/10.5555/3295222.3295349>
[Google Scholar](#) [WorldCat](#)
- [2] Avi Goldfarb, Bledi Taska, and Florenta Teodoridis. 2023. Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings. 52(1) *Research Policy* 104653. <https://www.sciencedirect.com/science/article/abs/pii/S0048733322001743>
[Google Scholar](#) [WorldCat](#)
- [3] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

- [4] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, 74–81.
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)
- [5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
[Google Scholar](#) [WorldCat](#)
- [6] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with Bert. *arXiv preprint arXiv:1904.09675*.
- [7] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.
- [8] Yuhui Zhang, Ruocheng Wang, and Zhengping Zhou. 2019. *Improving Neural Abstractive Summarization via Reinforcement Learning with Bertscore*. Unpublished technical report.
https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26632588.pdf
- [9] Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Reevaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*.
- [10] Franck Dernoncourt, Mohammad Ghassemi, and Walter Chang. 2018. *A repository of corpora for summarization*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.
<https://aclanthology.org/L18-1509/>
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)
- [11] Evan Sandhaus. 2008. *The New York Times Annotated Corpus*. Web Download.
<https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/GZC6PL>
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)
- [12] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. *An notated gigaword*. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction, AKBC-WEKEX '12*. Association for Computational Linguistics, USA, 95–100.
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)
- [13] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, 708–719.
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

- [14] Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, Arianna Bisazza and Omri Abend (eds), November. Association for Computational Linguistics, 300–309. <https://aclanthology.org/2021.conll-1.24>
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)
- [15] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. 12 *Transactions of the Association for Computational Linguistics* 39–57.
[Google Scholar](#) [WorldCat](#)
- [16] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.
- [17] Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.
- [18] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. 9 *Transactions of the Association for Computational Linguistics* 391–409.
[Google Scholar](#) [WorldCat](#)
- [19] Roei Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, 562–3591.
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)
- [20] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. *G-eval: NLG Evaluation Using GPT-4 with Better Human Alignment*, Singapore.
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)
- [21] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. 77 *Journal of Artificial Intelligence Research* 103–166.
[Google Scholar](#) [WorldCat](#)
- [22] Griffin Adams, Alexander Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. *arXiv preprint arXiv:2309.04269*.
- [23] Christian P. Janssen, Stella F. Donker, Duncan P. Brumby, and Andrew L. Kun. 2019. History and future of human-automation interaction. *International Journal of Human-Computer Studies* 99–107.
[Google Scholar](#) [WorldCat](#)
- [24] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. 267 *Artificial Intelligence* 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>

[Google Scholar](#) [WorldCat](#)

[25] Kartik Hosanagar. 2019. *A Human's Guide to Machine Intelligence*. Penguin Viking, New York.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 'Why Should I Trust You?' Explaining the predictions Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, 1135–1144.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

[27] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. *Interpretable decision sets: A joint framework for description and prediction*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, 1675–1684.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

[28] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*.

[29] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 1–16.

[30] Daehwan Ahn, Abdullah Almaatouq, Monisha Gulabani, and Kartik Hosanagar. 2024. Impact of model interpretability and outcome feedback on trust in AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, Association for Computing Machinery, Honolulu, HI, USA.

<https://doi.org/10.1145/3613904.3642780>

[31] Daniel Deutch, Nave Frost, and Amir Gilad. 2020. Explaining natural language query results. 29(1) *The VLDB Journal* 485–508.

[Google Scholar](#) [WorldCat](#)

[32] Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in NLP. 29(5) *Natural Language Engineering* 1199–1222.

[Google Scholar](#) [WorldCat](#)

[33] Marina Danilevsky, Shipi Dhanorkar, Yunyao Li, Lucian Popa, Kun Qian, and Anbang Xu. 2021. Explainability for natural language processing. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, 4033–4034, Virtual Event, Singapore.

<https://dl.acm.org/doi/10.1145/3447548.3470808>

[34] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Kraye, Francois Cadelon, and Karim R. Lakhani. 2023. *Navigating the Jagged*

Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper (24-013).

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

[35] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When scaling meets LLM finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*.

[36] Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I. Hong, and Adam Perer. 2023. *Zeno: An interactive framework for behavioral evaluation of machine learning*. In *CHI Conference on Human Factors in Computing Systems*, CHI '23, Association for Computing Machinery, New York, NY. <https://doi.org/10.1145/3544548.3581268>

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Notes

- 1 There are actually several variants of ROUGE, with the most widely used being ROUGE-N (which measures the overlap of n -grams between the generated summary and the reference summaries), ROUGE-L (which focuses on the longest common subsequence), and ROUGE-S (which considers skip-bigrams to capture sentence-level structure).
- 2 All the techniques discussed here are ‘referenced-based’ approaches, in that they depend on comparing a machine-generated summary to a reference human-written summary. While such techniques still require a significant amount of human labor to write or curate these reference summaries, the key value of these metrics is that they enabled the development of standardized, large-scale summarization datasets that the NLP community could benchmark against and iteratively build upon. With a high-signal and automatable method for comparing computer and human-generated summaries, a single human-written summary can be used in an unlimited number of experiments. Because of this, ROUGE unlocked the value of creating large datasets of text-summary pairs; this has led to the emergence of several summary-focused datasets in the NLP community [10,11,12,13]. These datasets have been crucial in the advancement of the field, as they allow scientists to evaluate many techniques and algorithms and efficiently compare their performance in a standardized way, iteratively identifying the strengths and weaknesses of different approaches through trial and error.
- 3 ‘Explanations increased team performance when the system was correct, but they decreased the accuracy on examples when the system was wrong, making the net improvement minimal’ (p. 14).
- 4 <<https://www.youtube.com/watch?v=3fl0b0eYxFQ>>.
- 5 Sample prompt: ‘If some comments express enthusiasm or a positive sentiment without much specific content about the product itself, include ‘General excitement’ or a similar theme in your summary. Outside of this, identify the specific topic of the commentary, such as hardware performance, hardware design, software performance, software design, service. Camera performance should be under either hardware or software depending on the nature of the feedback.’