Leveraging Large Language Models for Business Innovation: Hypothesis Generation in Product Development and Fraud Detection

by

Hassan Mohiuddin

S.B. in Computer Science and Engineering
Massachusetts Institute of Technology (2024)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

Authored by: Hassan Mohiuddin
Department of Electrical Engineering and Computer Science
December 11, 2024

Certified by: Amar Gupta
Research Scientist
Thesis Supervisor

Accepted by: Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Leveraging Large Language Models for Business Innovation: Hypothesis Generation in Product Development and Fraud Detection

by

Hassan Mohiuddin

Submitted to the Department of Electrical Engineering and Computer Science
on December 11, 2024, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

This thesis explores two key projects leveraging LLMs to enhance and augment human decision-making in product development processes. In our first project, we present a human-centered, two-phase approach to improving product innovation discussions. The initial phase introduces a hybrid methodology for summarizing significant amounts of customer reviews that combines SBERT embeddings with hierarchical clustering and LLMs. When tested on the Amazon reviews dataset, our approach achieves strong performance with BERT scores of 0.89 (precision), 0.88 (recall), and 0.86 (F1), while also performing well in human evaluations. Building on these condensed insights, the second phase develops a multi-agent LLM framework that facilitates professional product development discussions. This collaborative framework demonstrated strong performance in evaluations through Amazon Mechanical Turk, particularly excelling in its ability to simulate expert roles and generate relevant contributions. To explore potential real-world applications, we present a detailed case study of how this framework could be applied to banking fraud detection, outlining both possible technical advantages like simultaneous multi-perspective analysis and potential organizational benefits such as reduced coordination overhead and improved resource allocation.

In our second project, we tackle the optimization of A/B testing through a two-stage approach: first using LLMs to generate diverse headline candidates, followed by a validation methodology that supports human content creators in their decision-making process. Testing on the Upworthy dataset, our few-shot RAG-based system achieved 86% accuracy in predicting pairwise relative headline performance, demonstrating the potential for reducing dependency on traditional A/B testing while maintaining human oversight. Together, these projects demonstrate how LLMs can transform critical aspects of product development, from enhancing cross-functional team collaboration to streamlining content optimization decisions. While these implementations show promising results, we outline several areas for improvement and future

directions, establishing a foundation for continued research in human-AI collaborative frameworks.

Thesis Supervisor: Amar Gupta
Title: Research Scientist

Acknowledgments

I would like to extend my sincere thanks to Dr. Amar Gupta for his invaluable support and guidance throughout my MEng journey. His deep technical expertise not only helped shape the direction of my thesis but also challenged me to think critically about the real-world applications of my research. Beyond the laboratory, Dr. Gupta provided me with numerous opportunities for professional growth, from mentoring undergraduates to presenting our work to external organizations. The autonomy and initiative he encouraged helped me develop both as a researcher and as a professional. I am also grateful to my fellow lab members who supported and collaborated with me throughout this project.

I would like to thank my friends. The friends I made at MIT have been an incredible source of motivation and joy. They consistently pushed me to become a better version of myself while making my years here both enjoyable and unforgettable.

Finally, I owe my deepest gratitude to my family, whose unwavering support has been the foundation of my success. My older brothers have been constant role models, while my sister's jokes (though I may never admit it) have kept me smiling. My parents' sacrifices made it possible for me to study at MIT—my father's encouragement to push beyond my comfort zone and reach for greater heights, and my mother's endless care and attention to even the smallest details of my well-being. Without their love and support, I wouldn't be where I am today.

Biographical Sketch

Hassan Mohiuddin was born in Toronto, Canada, and completed his high school education at Brighton High School (BHS) in Rochester, New York. He enrolled at the Massachusetts Institute of Technology (MIT) in 2020, where he pursued a dual degree in Electrical Engineering & Computer Science and Management. During his undergraduate years, he conducted research at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) under the guidance of Dr. Amar Gupta, focusing on investigating the emerging applications of Large Language Models, particularly their impact in business and product development domains.

Beyond his academic pursuits, Hassan has cultivated diverse professional experience through internships across the technology and finance sectors. His work experience spans from innovative startups to established technology companies and quantitative hedge funds, where he has developed extensive expertise in software engineering and problem-solving. Building upon his undergraduate foundation, Hassan is continuing his academic journey at MIT, pursuing a Master of Engineering degree in Electrical Engineering and Computer Science while extending his research endeavors under Dr. Gupta's mentorship.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In a time characterized by swift technological progress, artificial intelligence (AI) has become a transformative force, altering industries and redefining how humans interact with machines. Among the many advancements in AI, Large Language Models (LLMs) are particularly noteworthy, representing a significant breakthrough in natural language processing (NLP) and paving the way for new opportunities in business innovation [10][54]. For example, in 1-1, we see that it took ChatGPT only 2 months to reach 100M users - significantly faster than other platforms. This thesis examines the potential of LLMs on product development and marketing strategies, providing insights into how these powerful tools can foster business growth and innovation in the digital era.

The evolution of AI, from its theoretical beginnings to its current advanced state, has been marked by several key breakthroughs. However, the emergence of LLMs signifies a fundamental shift in our understanding of machine learning and AI [17]. These models, which can comprehend, generate, and manipulate human language with remarkable accuracy and nuance have captivated researchers, business leaders, and the general public [43]. Their potential uses extend across a wide range of fields, including customer service, content creation, complex problem-solving, and decision support systems. Table 1.1 summarizes some primary differences between tradition NLP methods and LLMs.

At the core of this transformation is a rethinking of how machines process and

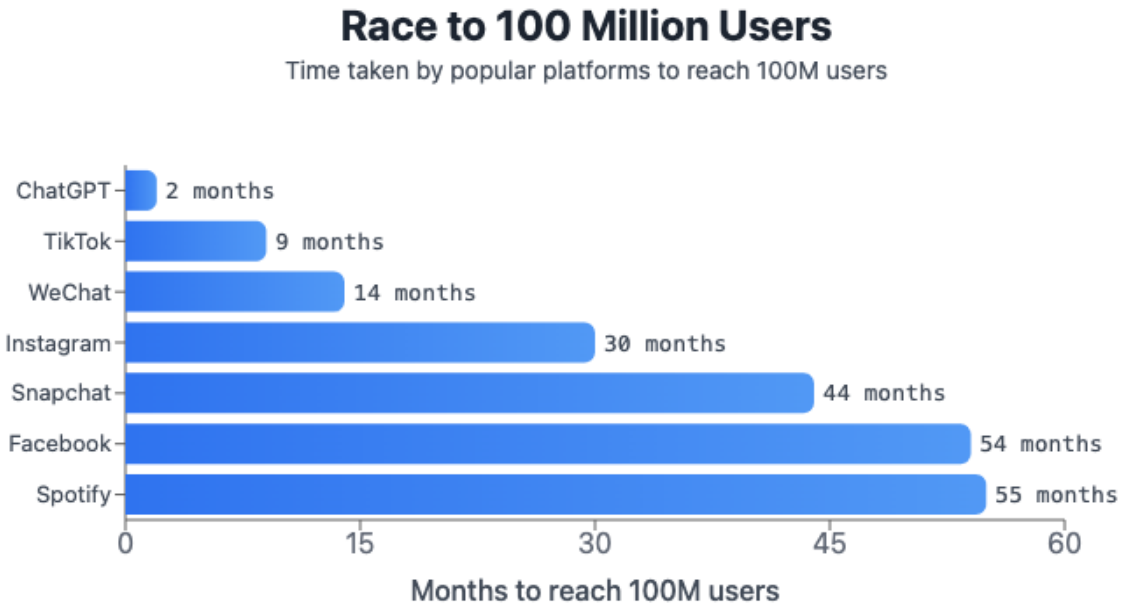

Figure 1-1: ChatGPT reached 100M users in just 2 months, significantly faster than any other platform [16].

interpret language. While traditional NLP methods have been useful, they often faced challenges with the complexity and ambiguity of human communication. In contrast, LLMs utilize extensive datasets and advanced neural network architectures to grasp the nuances of language in ways that more closely resemble human thought processes [10]. This significant advancement has been facilitated by a combination of factors, including improvements in computing power, access to large datasets, and innovative algorithmic approaches.

1.1 The Emergence of Transformer Architecture

The narrative of LLMs is incomplete without recognizing the crucial impact of the Transformer architecture. Introduced in 2017 by Vaswani et al. [54], the Transformer marked a departure from previous approaches to sequence processing in neural networks. Unlike recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, which process sequential data in order, the Transformer utilizes an attention mechanism that allows it to weigh the importance of different parts of the input dynamically [54].

Aspect	Traditional NLP Methods	Large Language Models
Context Understanding	Limited	Extensive
Task Adaptability	Task-specific	Multi-task capable
Training Data Required	Moderate	Massive
Inference Speed	Generally faster	Can be slower
Few-shot Learning	Limited	Strong capability
Interpretability	Often more interpretable	Less interpretable
Resource Requirements	Lower	Very high

Table 1.1: Comparison of Traditional NLP Methods vs. LLMs [44][65][66]

This innovation addressed several limitations of previous models:

1. **Parallelization:** By eliminating the need for sequential processing, Transformers can be trained much more efficiently on parallel computing architectures [54].

2. **Long-range dependencies:** The attention mechanism allows the model to capture relationships between words or tokens that are far apart in the sequence, a task that RNNs often struggled with [54].

3. **Contextual understanding:** Transformers can better capture the context in which words appear, leading to more nuanced and accurate language understanding [54].

The impact of the Transformer architecture on NLP is truly remarkable. It has enabled the development of models that can process and generate text with a level of coherence and contextual awareness that was previously unattainable. This breakthrough lays the groundwork for the next leap forward: the creation of LLMs.

1.2 The Evolution of Large Language Models

Building on the foundation established by the Transformer architecture, LLMs have become the forefront of NLP technology. These models, known for their vast scale and

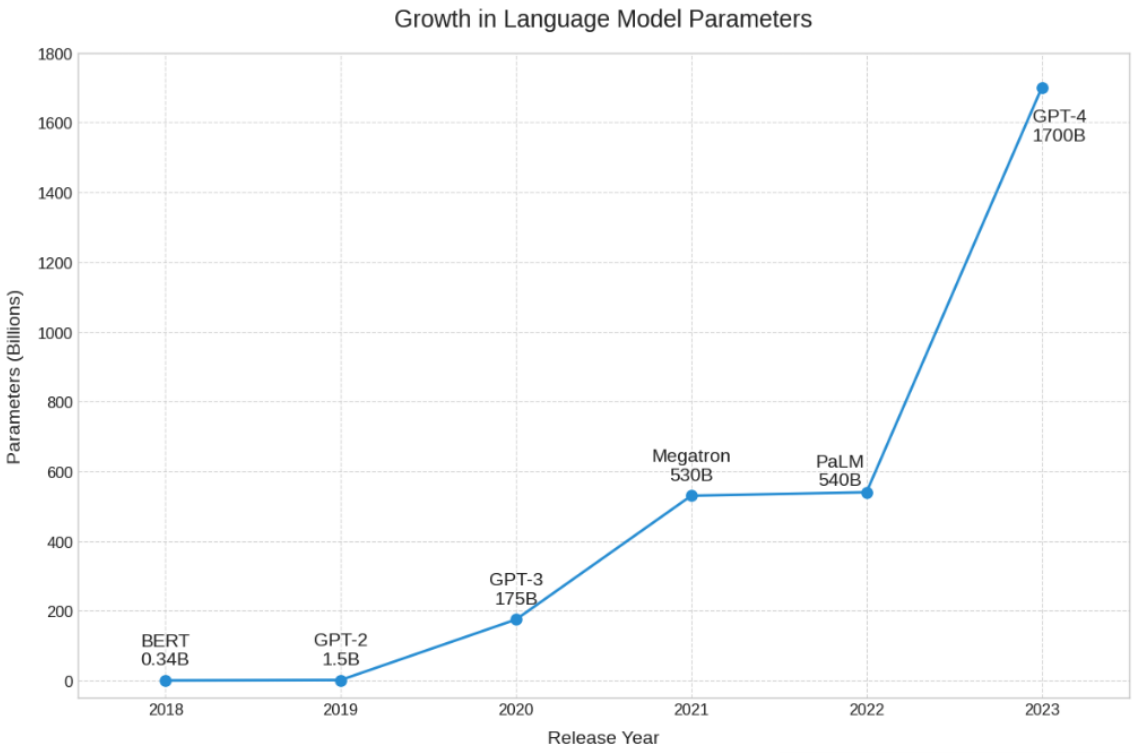

Figure 1-2: Evolution of NLP Model Sizes [15][39][50]

comprehensive training on a wide range of text sources, have shown abilities that make it difficult to distinguish between machine processing and human-like comprehension of language [10].

The evolution of LLMs has been marked by several key developments:

1. **Scale and Complexity:** Modern LLMs often encompass billions of parameters, allowing them to capture intricate patterns and relationships in language data. This scale enables them to perform a wide range of language tasks without task-specific training [10]. Figure 1-2 highlights the scaling of parameters in LLMs over the years.

2. **Pre-training and Fine-tuning:** The adoption of unsupervised pre-training on vast datasets, followed by task-specific fine-tuning, has significantly enhanced the versatility and performance of these models. This approach allows LLMs to develop a broad understanding of language before being adapted to specific applications [17].

3. **Generative Capabilities:** Unlike their predecessors, which were often limited to analysis or classification tasks, LLMs excel in generating coherent, contextually relevant text across a wide range of applications. This includes everything from creative writing to technical documentation [39].

4. **Few-shot and Zero-shot Learning:** Advanced LLMs have demonstrated the ability to perform new tasks with minimal or no specific training examples, a capability that closely mimics human cognitive flexibility [10].

5. **Multimodal Integration:** Recent developments have seen LLMs begin to integrate understanding across different modalities, such as text and images, further expanding their potential applications [39].

These advancements have culminated in models like GPT (Generative Pre-trained Transformer) series, BERT (Bidirectional Encoder Representations from Transformers), and their derivatives, which have set new benchmarks in language understanding and generation tasks.

1.3 Expanding Use Cases and Business Potential

The rapid advancement of LLMs has opened up a new world of possibilities across various sectors, particularly in business applications [38]. As organizations grapple with increasing data volumes and the need for more sophisticated analysis and decision-making tools, LLMs offer a powerful solution that can be applied to a wide range of challenges [8].

Key areas where LLMs are making significant impacts include:

1. **Advanced Data Analysis:** LLMs can process and synthesize vast amounts of unstructured data, uncovering insights that might otherwise remain hidden. This capability is particularly valuable in market research, competitive analysis, and trend forecasting [55].

2. **Enhanced Customer Interactions:** From sophisticated chatbots to personalized content generation, LLMs are revolutionizing how businesses communicated with their customers. They enable more natural, context-aware interactions that can significantly improve customer satisfaction and engagement [4].

3. **Product Development and Innovation:** By analyzing market trends, customer feedback, and technical documentation, LLMs can contribute to more informed and creative product development processes. They can generate hypotheses for product improvements, simulate customer reactions to potential features, and even assist in technical design processes [36].

4. **Content Creation and Marketing:** LLMs are transforming content marketing by generating high-quality, tailored content at scale. They can assist in creating everything from social media posts to long-form articles, adapting tone and style to specific audience segments [14].

5. **Decision Support:** The ability of LLMs to process complex information and generate human-like insights is proving invaluable in strategic decision-making. They can summarize vast amounts of information, generate scenarios, and even provide alternative viewpoints to challenge assumptions [11].

6. **Automation of Cognitive Tasks:** LLMs are enabling the automation of many cognitive tasks that previously required human intervention. This includes document summarization, translation, and even basic forms of creative work [19].

As businesses continue to explore and integrate LLM technologies, we are witnessing a paradigm shift in how organizations approach innovation, customer engagement, and operational efficiency. The potential for LLMs to augment human capabilities, rather than simply automate existing processes, opens up new avenues for value creation and competitive differentiation [9][10].

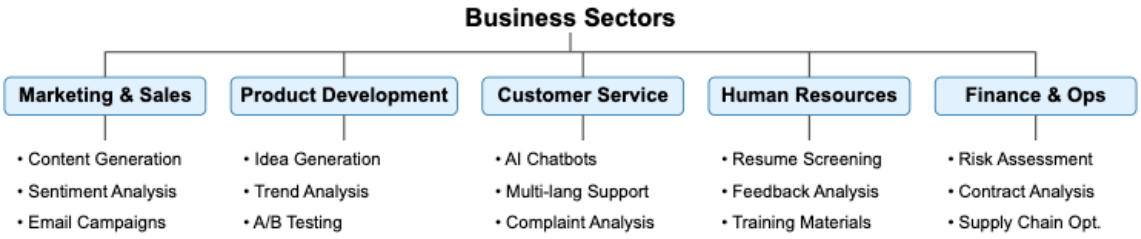

Figure 1-3: Business Sectors and Their AI Applications

However, the integration of LLMs into business processes is not without challenges. Issues such as bias in model outputs, the need for domain specific fine-tuning, and concerns about the interpretability and explainability of AI-generated insights must be carefully addressed [5]. Additionally, as LLMs become more deeply embedded in business operations, questions of ethics, governance, and responsible AI use come to the forefront [60].

This thesis aims to delve deeper into specific applications of LLMs in business contexts, focusing on their potential to drive product development and enhance marketing strategies. By exploring novel approaches to leveraging LLMs for hypothesis generation and A/B testing, we seek to contribute to the growing body of knowledge on practical applications of these powerful tools in business innovation [38].

1.4 Project Overview

The structure of this thesis is designed to provide a logical progression from foundational concepts to specific applications and experimental findings. In Chapter 1, we lay the groundwork with a comprehensive introduction to LLMs. This chapter traces the evolution of LLMs from their conceptual origins to their current state of sophistication, highlighting the key technological breakthroughs that have made them possible. We explore the fundamental principles underlying LLMs, their capabilities, and the broad impact they are having across various sectors, with a particular emphasis on their potential in business applications.

Chapter 2 presents a comprehensive review spanning four key domains foundational to our research. We begin by examining LLMs in customer feedback analy-

sis, tracking the evolution from traditional sentiment analysis to modern approaches that demonstrate superior performance in understanding customer satisfaction, particularly in e-commerce contexts. The chapter then explores hypothesis generation research, covering both scientific applications and creative problem-solving, with particular attention to how LLMs can potentially overcome human cognitive biases in ideation processes. The review continues with an analysis of LLM role-playing capabilities, from single-character simulation to sophisticated multi-agent frameworks that enable collaborative problem-solving. Finally, we examine A/B testing methodologies and content optimization techniques, including recent developments in LLM-driven experimental design. This literature review reveals a significant research opportunity in combining these elements for product innovation, particularly in leveraging multi-agent LLM systems for business decision-making—a gap our research aims to address.

Chapter 3 introduces a two-phase methodology for enhancing product innovation through LLMs, with the first phase focusing on extracting insights from customer feedback using the Amazon reviews dataset. Our approach combines SBERT embeddings with hierarchical clustering to group thematically similar reviews, following by LLM-powered abstractive summarization to distill key insights. Through both automated metrics and human evaluation, we demonstrate that this hybrid methodology accurately summarizes the reviews and identifies subtle product features that are often overlooked in conventional summaries.

Building on these condensed insights, the second phase presents a novel multi-agent LLM framework that simulates professional product development discussions. We detail our system's architecture, including carefully crafted role-specific prompts that enable different LLM instances to authentically represent various professional perspectives, from engineering to product management. Through expert interviews and case studies, industry professionals validated both the authenticity of the simulated interactions and the practical utility of the generated insights, particularly highlighting the system's ability to produce actionable recommendations while maintaining appropriate technical depth and business acumen.

Chapter 4 explores the intersection of LLMs and A/B testing, proposing a two-stage approach to optimize content testing in product development. While traditional A/B testing remains a cornerstone for validating design decisions, our methodology leverages the generative capabilities of LLMs to dramatically expand the testing landscape. The first stage uses LLMs to generate diverse headline variants, creating a rich pool of candidates for evaluation.

The second stage introduces a framework that is developed and tested using the Upworthy dataset [37], which contains real-world performance metrics including click-through rates and impression data. Our research investigates whether transformer-based machine learning models and LLMs can effectively predict user engagement metrics without the need for extensive live testing. We evaluate several architectural approaches, ranging from straightforward implementations using MAE and MSE loss functions to more sophisticated designs incorporating cross-attention mechanisms and targeted feature engineering. Additionally, we explore various LLM-based prediction strategies, including zero-shot and few-shot learning, enhanced reasoning techniques combining chain-of-thought and retrieval augmented generation, and a multi-LLM ensemble approach.

Chapter 2

Related Work

The integration of LLMs in business processes has led to significant advancements in understanding customer feedback and generating innovative product ideas. This section reviews existing research relevant to the key components of our study: customer feedback analysis, hypothesis generation for product development, review condensation techniques, advanced LLM workflows like multi-agent prompting, and literature around A/B testing.

2.1 LLMs in Customer Feedback Analysis

Customer feedback is a crucial source of insights for product improvement, yet traditional methods of analysis are often manual, time-consuming, and prone to bias.

Customer review analysis, a crucial component of product improvement, has traditionally relied on sentiment analysis and topic modeling techniques. Liu [34] provided a comprehensive survey of sentiment analysis methods, and Blei et al [7] introduced Latent Dirichlet Allocation for topic modeling. More recently, Zhang et al. [61] applied deep learning approaches to this domain.

The potential of LLMs to transform business processes has been increasingly recognized in recent years. Brown et al. [10] demonstrated the versatility of GPT-3 across various tasks including business applications. Moreover, there has been recent research specifically focused on utilizing LLMs with customer reviews. One study by

Table 2.1: Comparison of Key Papers and Their Contributions to Customer Feedback Analysis

Paper	**Key Contribution**	**Methodology**	**Results**	**Implications**
Our Work	Two-phase approach for enhancing product innovation discussions	SBERT embeddings, hierarchical clustering, LLM-powered summarization	BERT scores of 0.89 in precision, performs well in qualitative metrics	Provides richer, more comprehensive insights for product teams
Liu et al. [34]	Comprehensive survey of sentiment analysis methods	Review of various sentiment analysis techniques	N/A	Established foundation for modern sentiment analysis in product reviews
Zhang et al. [61]	Deep learning approaches for customer review analysis	Neural network models for sentiment analysis and feature extraction	N/A	Demonstrated potential of deep learning in understanding customer feedback
Azov et al. [4]	Automated system for replying to user reviews	Retrieval Augmented Generation (RAG)	Generated high-quality, human-like responses	Potential for scaling customer interaction and support
Falatouri et al. [20]	Extraction of sentiment and service quality dimensions from user content	LLM-based analysis	LLMs surpass prior AI models in both tasks	Highlighted need for human supervision to ensure reliability
Roumeliotis et al. [47]	Comparative analysis of LLMs for customer satisfaction	Evaluation of GPT-3.5, LLaMA-2, BERT, and RoBERTa	GPT-3.5 performed best, followed by LLaMA-2	Fine-tuned versions more effective than base versions

Azov et al. [4] automated the process of replying to user reviews. They created a system using Retrieval Augmented Generation (RAG) which generated high-quality, human like responses. Another study by Falatouri et al. [20] attempted to extract sentiment analysis and service quality dimensions from user-generated content. They found that LLMs surpass prior AI models in both tasks; however, human supervision is still necessary to ensure reliability.

Roumeliotis et al. [47] conducted a comparative analysis of GPT-3.5 and LLaMA-2 along with two additional Natural Language Processing (NLP) models, BERT and RoBERTa to determine if these models could contribute to understanding customer satisfaction within the context of an e-commerce environment. They found that GPT-3.5 performed the best followed by LLaMA-2, and the fine-tuned versions of each model were more effective than the base version. Our research builds on existing literature by proposing a methodology that combines more traditional machine learning techniques along with LLMs to process large-scale datasets.

2.2 Hypothesis Generation

2.2.1 Scientific Domains

The field of hypothesis generation using LLMs has gained attention in recent years, with researchers exploring various approaches to leverage the capabilities of these models in scientific discovery and creative problem-solving.

Several studies have focused on the application of LLMs in generating hypotheses for scientific domains. For instance, Zhou et al. [64] examine the potential of LLMs to generate hypotheses based on labeled examples. Their approach involves generating initial hypotheses from a small number of examples and then iteratively updating them to improve quality. Inspired by multi-armed bandits, they designed a reward function to inform the exploitation-exploration tradeoff in the update process. Their algorithm demonstrated significant improvements in predictive performance compared to few-shot prompting and even supervised learning in some cases.

In the biomedical field, Qi et al. [42] conducted a comprehensive evaluation of LLMs as biomedical hypothesis generators. They constructed a dataset of background-hypothesis pairs from biomedical literature and assessed the hypothesis generation capability of high-quality models in various settings. Their study incorporated tool use and multi-agent interactions to enhance the exploration of uncertainty, a crucial aspect of scientific discovery. They proposed novel metrics for evaluating the quality of generated hypotheses and found that LLMs can generate novel and validated hypotheses, even when tested on unseen literature.

2.2.2 Inductive Reasoning and Abstract Hypothesis Generation

Wang, et al. [56] explored the inductive reasoning capability of LLMs, focusing on complex tasks such as Abstraction and Reasoning Corpus (ARC). They proposed a method to improve LLMs' inductive reasoning by generating explicit hypotheses at multiple levels of abstraction. Their approach involves prompting the LLM to propose abstract hypotheses in natural language, then implementing these as concrete Python programs. This method showed improved performance on various inductive reasoning benchmarks, highlighting the importance of combining abstract hypothesis generation with concrete implementations.

2.2.3 Creativity and LLMs

While not directly focused on hypothesis generation, the work of Franceschelli et al. [22] on creativity in LLMs provides valuable insights into the potential of these models for generating novel ideas. They analyzed the development of LLMs through the lens of creativity theories, focusing on dimensions such as value, novelty, and surprise. The paper primarily focuses on whether LLMs can be labeled as "creative".

2.2.4 Human Cognitive Biases in Design

Jannson et al. [26] conducted a seminal study on design fixation, demonstrating how exposure to example solutions can constrain designers' ability to generate novel ideas. In their paper, the researchers showed that engineers who were shown an example solution to a design problem tended to incorporate more features from that example in their own designs, even when some of those features were problematic. This fixation effect persisted even when participants were explicitly instructed to avoid using features from the example. This study highlights a significant challenge in human idea generation: our tendency to fixate on existing solutions or familiar concepts, which can limit our ability to generate truly novel hypotheses or ideas. This is particularly relevant in complex problem-solving scenarios where innovative thinking is crucial. There is a potential for LLMs to serve as cognitive aids, supplementing human creativity and potentially overcoming cognitive biases or limitations.

2.3 Multi-LLM Frameworks

In recent years, the application of LLMs has extended beyond traditional natural language processing tasks to more complex simulations of human behavior and creativity. Notably, Character-LLM [49] introduced a method for training LLMs to embody specific historical figures through experience reconstruction and supervised fine-tuning, enabling sophisticated role-playing capabilities. This approach aligns with the broader categorization of LLM persona research outlined by [53], which distinguishes between LLM role-playing and LLM personalization. Within this framework, researchers have explored task-oriented dialogue modeling and user persona modeling [53] to enhance LLMs' ability to engage in goal-directed conversations and tailor responses to individual users.

Building on these foundations, recent work has focused on leveraging multi-LLM setups to induce collective creativity and problem-solving. The LLM Discussion Framework [35] represents a significant advancement in this area, proposing a struc-

Table 2.2: Comparison of Multi-Agent LLM Papers and Their Contributions

Paper	**Key Contribution**	**Methodology**	**Results**	**Implications**
Our Work	Multi-agent LLM framework for simulating product development discussions	Role-based LLM agents, cross-functional dialogue simulation	Qualitatively performs well in the areas of expert role simulation and practical utility	Potential to improve product team collaboration and decision-making
Jin et al. [29]	Comprehensive survey distinguishing between LLMs and LLM-based agents in software engineering	Systematic review of six key SE topics: requirements, code generation, autonomous decisions, design, testing, and maintenance	N/A	Demonstrated potential of AI collaboration in technical domains
Shao et al. [49]	Character-LLM for training LLMs to embody specific personas	Experience reconstruction and supervised fine-tuning	Enabled sophisticated role-playing capabilities	Opened possibilities for more nuanced and context-specific AI interactions
Lu et al. [35]	LLM Discussion Framework for creative ideation	Structured three-phase approach (initiation, discussion, convergence)	Superior performance in various creativity metrics compared to single-LLM setups	Demonstrated potential for collaborative problem-solving among AI agents

tured three-phase approach (initiation, discussion, convergence) to facilitate creative ideation among multiple LLM agents. This framework builds upon previous multi-LLM collaboration studies [18] that primarily focused on improving factuality and reasoning skills. By incorporating role-playing techniques and fostering diverse perspectives, the LLM Discussion approach has demonstrated superior performance in various creativity metrics compared to single-LLM setups [35].

These developments collectively underscore the evolving landscape of LLM applications, highlighting the potential for these models to not only simulate complex characters and personalized interactions but also to engage in collaborative, creative problem-solving. Building upon existing literature, our project extends the application of LLM-based character simulation and multi-agent creativity to the domain of business and product development, an area that has been largely unexplored in current research. While previous work has focused on historical figures [49] or general creativity tasks, we aim to leverage these techniques to simulate realistic business scenarios and enhance collaborative product development process.

2.4 A/B Testing

Content optimization through experimentation has become a crucial practice for digital platforms and publishers. This section reviews relevant literature on A/B testing methodologies, headline optimization techniques, and the emerging role of LLMs in content experiments.

2.4.1 Methodologies

A/B testing, also known as split testing, is a widely used method for comparing two or more variants of content to determine which performs better according to specified metrics. Larsen et al. [32] provide a comprehensive review of statistical challenges in online controlled experiments, highlighting issues such as sample ratio mismatch, network effects, and novelty effects that can impact the validity of A/B test results.

The dynamics of information spreading in social networks during A/B tests have

Table 2.3: Comparison of A/B Testing and Headline Optimization Papers

Paper	**Key Contribution**	**Methodology**	**Results**	**Implications**
Our Work	Two-stage methodology for optimizing A/B testing	LLM-generated candidates, few-shot RAG-based validation	86% accuracy in predicting pairwise relative headline performance	Potential to significantly reduce dependency on traditional A/B testing methods
Ye et al. (2024)	LOLA (LLM-Assisted Online Learning Algorithm)	Combination of LLM prediction models with adaptive experimentation	Superior performance compared to traditional A/B testing and pure online learning algorithms	Particularly effective in scenarios with limited experimental traffic or numerous content variants
Jin et al. (2020)	Novel system for generating stylized headlines beyond factual summarization	Multitasking framework combining summarization and reconstruction tasks with parameter sharing for style disentanglement	Surpassed state-of-the-art summarization models by 9.68% in attraction scores, outperformed human-written references	Demonstrated feasibility of generating engaging, style-specific headlines without style-specific training data, opening new possibilities for content optimization
Kuiken et al. (2017)	Study of effective newspaper headlines in digital environments	Analysis of headline characteristics	Identified key features of effective headlines (e.g., short words, signal words, sentimental words)	Provided insights for optimizing headline engagement in digital media

been explored by Ottaviani et al. [40]. Their work simulates A/B testing scenarios to determine effects on the dissemination of information through networks, employing LASSO regression to identify overall impacts on click-through rates. This research underscores the importance of considering network effects when interpreting A/B test results in social media contexts.

To address the limitations of traditional A/B testing, several advanced methodologies have been proposed. Chennu et al. [13] introduce methods to maximize the extraction of learnings from past A/B tests, enabling more efficient use of historical data. Jeunen et al. [27] suggest ways to maximize decisions made from results to improve A/B testing for long-term cases and reduce the sample sizes needed for statistically significant outcomes.

Offline A/B testing has emerged as a valuable approach for evaluating new technologies without live user experimentation. Wu [57] proposes AutoOffAB, a framework for automated offline A/B testing in data-driven requirement engineering. They highlight the potential for LLM-based approaches to create more intelligent non-trivial variants for offline A/B testing, extending capabilities beyond simple hyperparameter tuning.

2.4.2 Headline Optimization

Headline optimization is a critical aspect of content strategy for digital publishers. Several studies have focused on generating effective and attractive headlines using various techniques. Jin et al. [28] developed a headline generation system capable of producing headlines with controlled styles, such as funny, romantic, or clickbait. Their approach leverages datasets of stylized headlines and employs both human evaluation (based on relevance, attractiveness, fluency, and style strength) and automatic evaluation (using summarization quality metrics and language fluency measures).

Song et al. [51] proposed a framework using CNN and LSTM architectures to generate eye-catching headlines. Their user study evaluated generated headlines based on attractiveness, relevance, and grammaticality, with 63% of participants affirming that the generated headlines were more attractive than original ones.

Xu et al. [58] introduced a reinforcement learning approach for generating sensational headlines, balancing reinforcement learning with maximum likelihood estimation. Their model achieved a 60% sensationalism rate in human evaluations.

In a study of effective newspaper headlines in digital environments, Kuiken et al. [31] identified several characteristics of effective headlines, including the use of short words, absence of questions and quotes, inclusion of signal words and pronouns, and the presence of sentimental words.

The detection and measurement of clickbait have also been subjects of research. Potthast et al. [41] organized the Clickbait Challenge, which resulted in the creation of the Webis Clickbait Corpus and various approaches for clickbait detection and strength regression.

2.4.3 LLMs in Content Experiments

The application LLMs in content optimization and A/B testing is an emerging field with significant potential. Ye et al. [59] introduce LOLA (LLM-Assisted Online Learning Algorithm), a novel framework that combines the predictive power of LLMs with adaptive experimentation techniques to optimize content delivery. Their approach leverages LLM-based prediction models as priors in online learning algorithms, demonstrating superior performance compared to traditional A/B testing and pure online learning algorithms, particularly in scenarios with limited experimental traffic or numerous content variants.

The use of LLMs for sentiment analysis and misinformation detection in headlines has also been explored. Juroš et al. [30] investigate the application of LLMs for targeted sentiment analysis in news headlines, addressing the descriptive-prescriptive dilemma in this context. Similarly, Rony et al. [46] examine the potential of LLMs in identifying misleading news headlines, contributing to the broader effort of combating misinformation in digital media.

While existing research has explored various aspects of headline optimization and the use of LLMs in content experiments, our current study leverages LLMs to generate headlines specifically designed to increase click-through rates. This research not

only utilizes the generative capabilities of LLMs but also capitalizes on their ability to reflect average human sentiment and preferences. Recent studies have shown that LLMs can effectively replicate human judgments on tasks involving subjective labels derived from human experiences. For instance, works by Aher et al. and Santurkar et al. [2] [48] have demonstrated LLMs' capacity to mirror human responses in areas such as social science and psychological studies and opinion polling. By harnessing this capability, the present research aims to create headlines that are more likely to resonate with readers and drive engagement. This approach bridges the gap between automated content generation and human-centric design, potentially offering a more efficient and effective method for headline optimization in digital publishing. Moreover, we propose a validation methodology which aims to predict the best performing headlines among candidates.

Chapter 3

Novel Hypothesis Generation

In today's market, understanding and responding to customer feedback is crucial for continuous product improvement and innovation. Traditional methods of analyzing customer reviews often involve manual processes that are time-consuming and prone to bias. Our research builds on the advancements of LLMs and exploring how they can be used not just to analyze reviews, but to generate novel hypotheses for product improvements. This involves condensing large amounts of feedback into manageable insights and using advanced prompting techniques to facilitate creative and actionable idea generation.

3.1 Amazon Customer Reviews Dataset

3.1.1 Overview

To improve hypothesis generation, we utilize the Amazon US Customer Reviews data. This extensive collection of customer feedback, compiled over more than two decades since 1995, represents a comprehensive repository of consumer opinions and experiences regarding a vast array of products available on the Amazon.com website. For this thesis, we will focus on the software products.

The dataset offers extensive information for each review entry, summarized in 3.1. The relevance of this dataset to novel hypothesis generation for product improvement

is multifaceted. Firstly, the diversity of perspectives represented in the millions of reviews provides a solid base for innovative thinking. Unlike controlled studies, this dataset offers authentic, unsolicited customer opinions, capturing nuances that might be overlooked in more structured feedback mechanisms.

For our purposes, the combination of quantitative ratings with qualitative review text provides a holistic view of customer sentiment. This dual nature allows for both statistical analysis and in-depth qualitative exploration, enabling researchers to corroborate findings across different analytical approaches. The additional context provided by metadata such as verified purchases and helpful votes adds layers of credibility and relevance to the reviews, helping to prioritize most valuable feedback in the hypothesis generation process.

We utilize the textual contents to prompt LLMs and determine if these comments/reviews by product users can help create valuable suggestions and improvements for the product. One potential limitation of this dataset is the inherent self-selection bias, as reviews are voluntarily submitted. This could potentially over-represent extremely satisfied or dissatisfied customers. The dataset is also specific to Amazon, which may not fully represent the broader market perspective. Also, older reviews may not reflect current market conditions or product iterations, necessitating careful consideration of temporal context. Moreover, despite verification measures, there's always a possibility of manipulated or inauthentic reviews, which should be considered. Still, this dataset is a good starting ground for our experimentation and analysis.

3.1.2 Exploratory Data Analysis

To gain a comprehensive understanding of the Amazon Reviews Dataset for software products, we conduct an exploratory data analysis (EDA). This analysis focused on several key aspects of the data, including the distribution of average ratings, discrete ratings, number of reviews per product, and the relationship between ratings and helpful votes. These analyses provide valuable insights into user behavior, product reception, and the overall landscape of software product reviews on Amazon.

Table 3.1: Summary of Attributes in the Amazon US Customers Review Dataset

Attribute	Description
marketplace	2-letter country code of the review's marketplace
customer_id	Random identifier for aggregating reviews by author
review_id	Unique ID of the review
product_id	Unique ID of the product being reviewed
product_parent	Random identifier for aggregating reviews of the same product
product_title	Title of the product
product_category	Broad category for grouping products and reviews
star_rating	1-5 star rating of the review
helpful_votes	Number of helpful votes received
total_votes	Total number of votes received
vine	Indicates if the review was part of the Vine program
verified_purchase	Indicates if the review is on a verified purchase
review_headline	Title of the review
review_body	Text of the review
review_date	Date the review was written

3.1.2.1 Distribution of Average Ratings

We begin our analysis by examining the distribution of average ratings for software products. This gives us an overview of how products are generally perceived by users.

Figure 3-1 and Table 3-2 present the distribution and summary statistics of average ratings for software products. The mean rating of 3.3356 and median of 3.4 indicate that, on average, software products receive moderately positive reviews.

The standard deviation of 0.812 suggests a reasonable spread in the ratings, with most products falling between 2.5 and 4.2 stars. The negative skewness of -0.601 indicates that the distribution is slightly left-skewed, meaning there are more products with above average ratings than below-average ratings.

This could suggest a general satisfaction with software products or a potential bias towards positive reviews. The kurtosis of 1.248 indicates that the distribution has slightly heavier tails and higher peak compared to a normal distribution. This suggests that there are more extreme ratings (both very high and very low) than would be expected in a perfectly normal distribution.

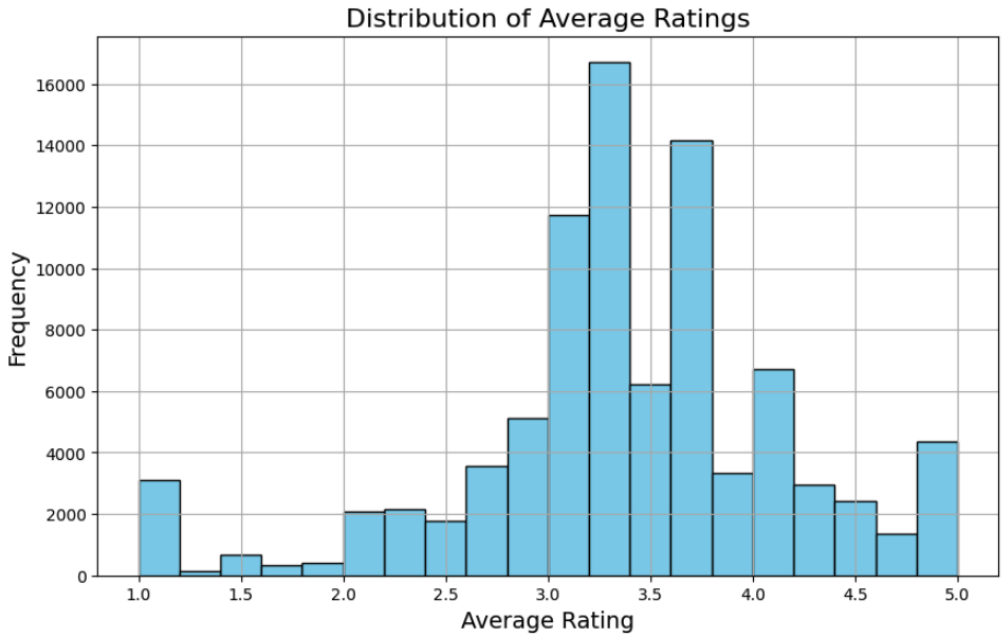

Figure 3-1: Distribution of Average Ratings for Software Products

Statistic	Value
Mean	3.3556
Median	3.4000
Standard Deviation	0.8130
Skewness	-0.6010
Kurtosis	1.2480

Figure 3-2: Summary Statistics for Average Ratings

3.1.2.2 Distribution of Discrete Ratings

To further understand the rating patterns, we analyze the distribution of discrete ratings (1 to 5 stars).

Figure 3-3 and Table 3-4 show the distribution of discrete ratings. The mean rating of 3.94 and median of 5.00 indicate a strong positive skew in the ratings. This is further confirmed by the negative skewness value of -1.10, which suggests that there are more high ratings than low ratings.

The standard deviation of 1.45 indicates a significant spread in the ratings, while the negative kurtosis (-0.29) suggests that the distribution is slightly flatter than a normal distribution, with fewer extreme values.

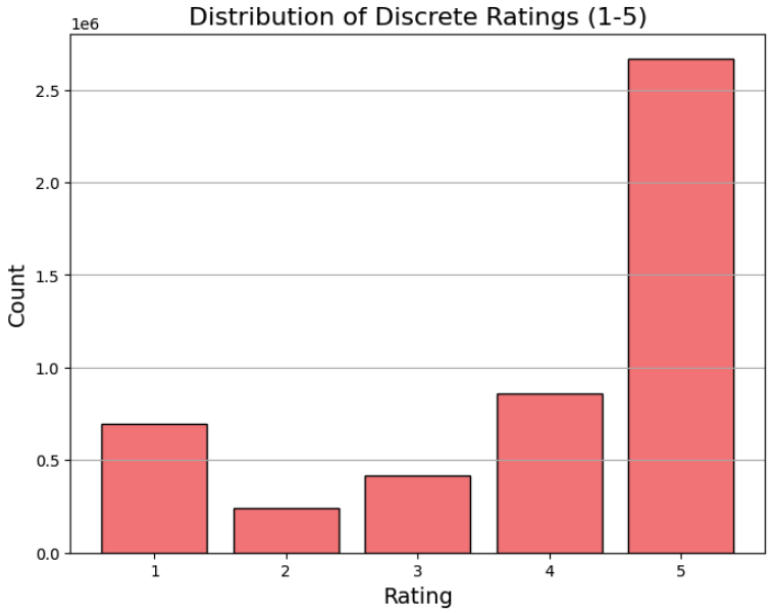

Figure 3-3: Distribution of Discrete Ratings for Software Products

Statistic	Value
Mean	3.94
Median	5.00
Standard Deviation	1.45
Skewness	-1.10
Kurtosis	-0.29

Figure 3-4: Summary Statistics for Discrete Ratings

This analysis reveals a tendency for users to give either very high (4 or 5 stars) or very low (1-star) ratings, with a clear preference for positive ratings. This polarization could be due to various factors, such as satisfied users being more likely to leave reviews, or the nature of software products leading to more extreme user experiences.

3.1.2.3 Distribution of Reviews per Product

Understanding the number of reviews per product helps us gauge the amount of user feedback available for different software items.

Figure 3-5 and Table 3-6 present the distribution of the number of reviews per product. The distribution is heavily right-skewed, as indicated by the large positive skewness (37.84) and the substantial difference between the mean (54.01) and median

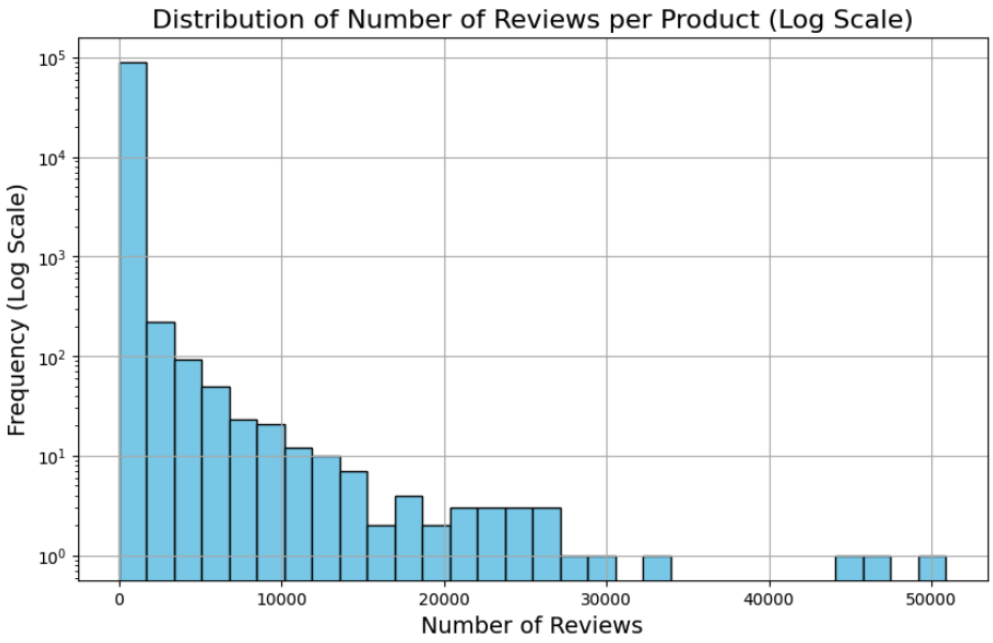

Figure 3-5: Distribution of Number of Reviews per Product

Statistic	**Value**
Mean	54.01
Median	3.00
Standard Deviation	593.91
Skewness	37.84
Kurtosis	2108.30

Figure 3-6: Summary Statistics for Number of Reviews per Product

(3.00) values.

The extremely high kurtosis (2108.30) suggests that the distribution has very heavy tails, indicating the presence of products with an exceptionally high number of reviews. The large standard deviation (593.91) compared to the mean further emphasizes the high variability in the number of reviews across products.

This distribution reveals that while most software products have relatively few reviews (as indicated by the low median), there are some products that attract an enormous number of reviews. This could be due to factors such as popularity, controversy, or longevity of the product in the market.

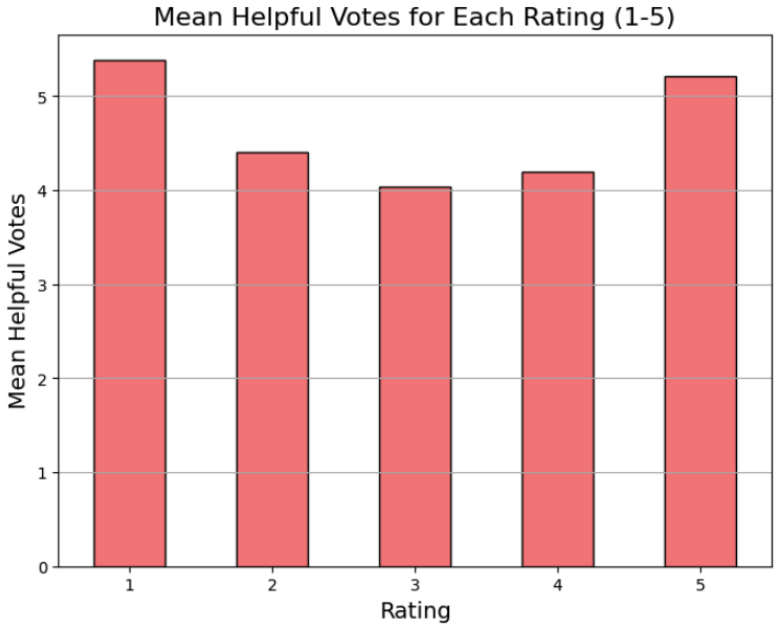

Figure 3-7: Distribution of Average Helpful Votes by Rating

Statistic	Value
Mean	4.65
Median	4.40
Standard Deviation	0.61
Skewness	0.31
Kurtosis	-1.71

Figure 3-8: Summary Statistics for Average Helpful Votes by Rating

3.1.2.4 Distribution of Helpful Votes by Rating

Finally, we examine the relationship between ratings and the average number of helpful votes received.

Figure 3-7 and Table 3-8 show the distribution of average helpful votes across different ratings. The mean (4.65) and median (4.40) are relatively close, indicating a fairly symmetric distribution. This is supported by the low skewness value (0.31).

The negative kurtosis (-1.71) suggests that the distribution is flatter than a normal distribution, indicating a relatively even spread of helpful votes across different ratings. The standard deviation (0.61) is relatively low compared to the mean, suggesting that the average number of helpful votes doesn't vary dramatically across

different ratings.

This analysis reveals that users find reviews across all rating levels to be helpful, with a slight preference for reviews at certain rating levels. This could indicate that users value both positive and negative reviews when making decisions about software products.

3.1.2.5 Discussion

The exploratory data analysis of the Amazon Reviews Dataset reveals important patterns in how users engage with and evaluate software products on the platform. While the analysis shows a generally positive reception of software products with an average rating of 3.34, the underlying patterns suggest a more complex landscape of user experiences and reviewing behavior.

A particularly interesting finding is the polarization in user ratings, with a notable preference for extreme ratings, especially five-star reviews. This pattern differs from what might be expected in a normal distribution and could be intrinsic to the nature of software products. Unlike physical goods, software products often have binary success conditions - either they work as intended for a user's specific setup and needs, or they don't. This characteristic might explain why users tend toward either very positive or very negative reviews, with fewer moderate opinions. The highly skewed distribution of review volumes across products presents both opportunities and challenges for analysis. The presence of products with exceptionally high review counts (as indicated by the kurtosis of 2108.30) suggests that certain software products generate significantly more user engagement than others. These outlier products merit further investigation, as they might offer valuable insights into what drives user engagement in the software market. However, this skewed distribution also raises questions about the representativeness of the review data for products with fewer reviews.

The analysis of helpful votes adds another dimension to our understanding of user behavior. The relatively uniform distribution of helpful votes across rating levels challenges the assumption that users might favor either positive or negative reviews.

Instead, it suggests that users value well-articulated opinions regardless of their valence, indicating sophisticated consumer behavior in evaluating software products. This finding has implications for how we might weight and prioritize reviews in future analyses.

These patterns collectively point to several areas worthy of future investigation. Temporal analysis could reveal how software products evolve and how user expectations change over time. Additionally, deeper content analysis of reviews, particularly those receiving high helpful votes across different rating levels, could uncover specific product attributes that drive user satisfaction or dissatisfaction. Such insights would be valuable for both software developers and marketplace platforms in improving product quality and user experience.

The findings also suggest potential methodological considerations for future research. The strong skew in review volumes indicates that sampling strategies should carefully account for products with varying levels of user engagement. Furthermore, the polarization in ratings suggests that analysis methods should be robust to non-normal distributions and extreme values.

3.2 Review Condensation

One challenge we face is the sheer volume of reviews for popular products on Amazon often exceeds the input capacity of LLMs. Moreover, it costs more resources to run inference on larger input sequences. This necessitates an effective review condensation strategy to distill the essence of customer feedback into a manageable and representative summary. Our approach to review condensation aims to create a comprehensive yet concise textual summary that captures the diverse perspectives present across different reviews. We propose a baseline and two distinct methods for review condensation, each offering unique insights into customer opinions.

Algorithm 1 Split Text Into Chunks

```
function SplitIntoChunks(text, maxLen)
    chunks ← []
    currentPos ← 0
    while currentPos < |text| do
        chunk ← text[currentPos : currentPos + maxLen]
        chunks.append(chunk)
        currentPos ← currentPos + maxLen
    end while
    return chunks
end function
```

Algorithm 2 Iterative Chunked Summarization

```
function IterativeSummarize(reviewText, modelContextLen, model)
    segments ← SplitIntoChunks(reviewText, modelContextLen)
    summaries ← []
    while |segments| > 1 do
        for chunk ∈ segments do
            chunkSummary ← model(chunk)
            summaries.append(chunkSummary)
        end for
        segments ← SplitIntoChunks(summaries, modelContextLen)
        summaries ← []
    end while
    return segments[0]
end function
```

3.2.1 Baseline

In developing our baseline approach, we opt for a straightforward solution using a BERT transformer model specialized in summarization. However, we face a challenge: our product reviews are longer than what the model's context length would allow. To address this, we break each review into smaller, manageable segments that fit within the model's context window. We follow a two-step process:

1. First, we generate summaries for each individual segment
2. Then, we combine these segment summaries and run them through the model until we create one cohesive final summary

This approach as described in 1 and 2, while simple, is well-established in the field and provides a good benchmark to evaluate the effectiveness of our other methods.

3.2.2 SBERT Based Methods

3.2.2.1 Sentiment-Based Review Segmentation

The first step in our methodology involves segmenting reviews based on sentiment. We use the star rating as a proxy for sentiment analysis:

- Negative sentiment: Reviews with 3 stars or below
- Positive sentiment: Reviews with 4 stars or above

This initial segmentation allows us to process positive and negative feedback separately, ensuring a balanced representation in our final analysis. Below, we propose two distinct methods for condensing the segmented reviews.

3.2.2.2 Top-N Most Helpful Reviews

This method selects the top N reviews for each star rating based on the number of helpful votes. The rationale behind this approach is that reviews with many helpful votes are likely to capture the general consensus and provide valuable insights. By

selecting top reviews from each rating category, we ensure a diverse range of perspectives.

3.2.2.3 Embedding-based Clustering and Sampling

This more sophisticated approach utilizes Sentence-BERT (SBERT) [45] to generate embeddings for each review, followed by clustering and representative sampling:

1. Generate SBERT embeddings for each review
2. Perform K-means clustering on the embeddings
3. Select representative samples from each cluster using a weighted approach

The selection of representative samples is based on a combination of similarity to the cluster centroid and the number of helpful votes. This ensures that we capture both the central themes of each cluster and the reviews that users found most helpful. The pseudocode can be seen in 3 and 4.

Algorithm 3 Select Examples From Cluster

function SELECTEXAMPLESFROMCLUSTER($clusterEmb, helpfulVotes, n, alpha$)
    $centroid \leftarrow \text{Mean}(clusterEmb)$
    $similarities \leftarrow \text{CosineSimilarity}(clusterEmb, centroid)$
    $normVotes \leftarrow helpfulVotes / \max(helpfulVotes)$
    $scores \leftarrow alpha \times similarities + (1 - alpha) \times normVotes$
    $topIndices \leftarrow \text{ArgSort}(scores)[-nEx :]$
    return $\text{Reverse}(topIndices)$
end function

3.2.2.4 Pros and Cons Extraction

Once we have representative reviews, we use an LLM to extract pros and cons based on the sentiment group:

- For negative sentiment reviews, we extract cons
- For positive sentiment reviews, we extract pros

We use the following prompts for the LLM:

Algorithm 4 Cluster and Select Representative Reviews

function ClusterAndSelectExamples($emb, votes, nClusters, n, alpha$)
    $labels \leftarrow$ ClusterReviews($emb, nClusters$)
    $selected \leftarrow$
    for $cId \leftarrow 0$ **to** $nClusters - 1$ **do**
        $cIndices \leftarrow$ Where($labels = cId$)
        $cEmb \leftarrow emb[cIndices]$
        $cVotes \leftarrow votes[cIndices]$
        $sIndices \leftarrow$ SelectExamplesFromCluster($cEmb, cVotes, nEx, alpha$)
        $gIndices \leftarrow cIndices[sIndices]$
        $selected[cId] \leftarrow gIndices$
    end for
    return *selected*
end function

Prompt 1: Generating Cons

```
You are an expert Amazon product reviewer. Your task is to analyze the
given product reviews and generate a list of cons (negative aspects)
of the product. Format the output as a JSON list of strings, each
representing a distinct con. Be specific.

Example input: "While the recipes taste great, many of them are quite
rich and calorie-heavy. It's not really a cookbook for those looking
for health-focused meals."

Example output: ["Recipes are calorie-dense", "Not focused on healthy
eating"]
```

Prompt 2: Generating Pros

```
You are an expert Amazon product reviewer. Your task is to analyze the
given product reviews and generate a list of pros (positive aspects)
of the product. Format the output as a JSON list of strings, each
representing a distinct pro. Be specific.

Example input: "This cookbook has so many delicious recipes! I love
how it's divided into different sections. Even my non-vegan friends
```

```
enjoy the meals I make from it."

Example output: ["Includes delicious recipes", "Well-organized into
sections", "Appealing to non-vegans"]
```

3.2.2.5 Comprehensive Review Summary

Finally, we use another LLM prompt to synthesize the pros and cons into a comprehensive review summary. This summary provides a balanced overview of the product, considering the full scope of customer opinions. Here's the prompt used for this task:

Prompt 3: Comprehensive Review Summary

```
You are an expert Amazon product reviewer. Your task is to analyze the
given pros and cons of a product and generate a comprehensive review
summary. Your output should be in JSON format with the following
structure:
 {
 "verdict": "A summary of the overall product impression, focusing
 on functionality and benefits. It can be a detailed
 summary.",
 "pros": "Top 2-5 most important pros, summarized and rephrased
 if necessary",
 "cons": "Top 2-5 most important cons, summarized and rephrased
 if necessary"
 }
Ensure that the verdict captures the essence of the product based on
the balance of pros and cons, focusing on functionality and potential
benefits. The summarized pros and cons should highlight the most
impactful aspects of the product.

Here's an example:

Input:
```

```
Pros: ["Excellent sound quality", "Comfortable fit", "Long battery
life", "Easy pairing process", "Water-resistant", "Good noise
cancellation", "Affordable price"]
Cons: ["Occasional connectivity issues", "App can be buggy", "Limited
touch controls", "Bulky charging case"]
Product Title: Wireless Earbuds Average Rating: 4.3

Output:
 {
 "verdict": "These wireless earbuds offer impressive sound quality
 and comfort with long battery life, making them an
 excellent value for everyday use.",
 "pros": [
 "Excellent sound quality",
 "Comfortable fit for extended wear",
 "Long-lasting battery life",
 "Easy and quick pairing process",
 "Water-resistant design for versatile use"
 ],
 "cons": [
 "Occasional connectivity issues reported",
 "Companion app may have some bugs"
 ]
 }

Now, please analyze the provided pros and cons and generate a similar
summary.
```

This comprehensive summary serves as a concise yet informative representation of the product reviews, which can be used in subsequent stages of analysis or for generating improvement suggestions using LLM agents.

3.2.3 Evaluation

For our evaluation, we utilized the AmaSum dataset, which represents a large abstractive opinion summarization dataset. AmaSum comprises over 33,000 human-written summaries of Amazon product reviews, with each summary corresponding to an average of 320 customer reviews. The summaries are structured to include verdicts, pros, and cons providing a comprehensive overview of product opinions.

We employed multiple evaluation metrics to assess the performance of our summarization methods.

3.2.3.1 ROUGE Scores

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [33] metrics are widely used for evaluating text summarization:

- **ROUGE-1**: Measures the overlap of unigrams between the generated summary and reference summary.
- **ROUGE-2**: Evaluates the overlap of bigrams, capturing more complex phrasal information.
- **ROUGE-L**: Considers the longest common subsequence between the generated and reference summaries, allowing for non-consecutive matches while maintaining sequence order.

3.2.3.2 BERTScore

BERTScore [62] leverages the pre-trained BERT embeddings to compute similarity scores by cosine similarity. It provides three key metrics:

- **Precision**: Measures how many tokens in the generated summary are semantically similar to the reference.
- **Recall**: Evaluates how many tokens in the reference summary are captured in the generated summary.

- **F1**: The harmonic mean of precision and recall.

3.2.3.3 LLM as Judge

We employed a LLM to evaluate summaries based on five key aspects [63]:

- **Accuracy**: Correctness of information
- **Completeness**: Coverage of key points
- **Conciseness**: Efficient information presentation
- **Coherence**: Logical flow and readability
- **Relevance**: Focus on important aspects

The use of LLMs as judges is appropriate for this task because they can understand semantic similarities beyond surface level lexical matching. They are able to evaluate multiple aspects of summary quality and maintain consistency in evaluation.

3.2.3.4 Results & Analysis

Table 3.2: Comparison of Summarization Methods

Method	ROUGE			BERTScore			LLM
	R-1	R-2	R-L	P	R	F1	
Baseline	0.15	0.02	0.11	0.75	0.73	0.74	0.45
SBERT+Top N	0.26	0.03	0.17	0.84	0.82	0.80	0.65
SBERT+HC	**0.30**	**0.05**	**0.19**	**0.89**	**0.88**	**0.86**	**0.70**

It's crucial to first note the fundamental difference between abstractive and extractive summarization approaches. Extractive summarization selects and arranges existing text fragments, while abstractive summarization generates new text that captures the essence of the source material. This distinction is particularly relevant when interpreting our results.

The relatively low ROUGE scores across all methods can be attributed to several factors. Primarly, ROUGE metrics capture lexical overlap, which may not effectively

evaluate abstractive summaries. Moreover, the inherent variability in human-written summaries means multiple valid summaries can exist for the same input. Abstractive summarization may also use different vocabulary while maintaining semantic meaning.

In contrast, BERTScore and LLM evaluations show more promising results, particularly for our SBERT-based methods. This aligns with expectations as these metrics capture better semantic similarity rather than strict lexical matching. The SBERT + HC method achieved the highest performance across all semantic metrics (in bold).

To complement our automated metrics and provide quantitative support for our qualitative observations, we conducted a human evaluation study focusing on the pros and cons generated by each method. For each product in our test set, workers were presented with two lists: the ground truth pros/cons from the AmaSum dataset and the generated pros/cons from our system. They were instructed to identify matching points between the lists, where a match was defined as points expressing the same core idea, regardless of specific wording.

Table 3.3 presents these results, which align with and strengthen our qualitative observations. The SBERT+HC method achieved the highest performance with an overlap of 67% for pros and 65% for cons, indicating strong coverage of ground truth points. The relatively high precision scores (64% for pros, 61% for cons) suggest that the generated points are largely valid and relevant.

Our qualitative analysis revealed several important insights beyond these metrics. While the overlap scores are significant, the generated summaries frequently identified valid points absent from ground truth summaries. This suggest our algorithmic approach may be more thorough in processing the review corpus, capturing perspectives that human summarizers might overlook. The variations between generated and reference summaries often represented equally valid alternative viewpoints rather than errors, highlighting both the subjective nature of summary generation and our method's ability to capture diverse perspectives.

The evaluation of abstractive summarization systems presents unique challenges, primarily due to the inherent subjectivity in defining what constitutes a "good" sum-

mary. Multiple valid ways to present the same information make it difficult to establish a single "correct" reference, as reflected in the overlap and precision metrics. While automated metrics like ROUGE and BERTScore provide valuable insights, they may struggle to capture nuanced aspects of language and meaning, potentially missing subtle semantic relationships or failing to recognize valid alternative phrasings. Furthermore, reference summaries may contain biases from human writers who prioritize certain aspects based on personal experience.

Despite these challenges, our combined approach of automated metrics with human evaluation demonstrates the effectiveness of the SBERT+HC method. The strong performance across both automated metrics (Table 3.2) and human evaluation (Table 3.3) supports our conclusion that the method generates meaningful, comprehensive summaries that maintain coherence while capturing the essence of source reviews. Notably, our approach demonstrates potential for reducing human bias by systematically analyzing all available reviews, ensuring minority opinions and less prominent points are not overlooked.

Table 3.3: Human Evaluation of Pros/Cons Generation

Method	**Pros**		**Cons**	
	Overlap	**Precision**	**Overlap**	**Precision**
Baseline	0.45	0.38	0.41	0.35
SBERT+Top N	0.58	0.54	0.55	0.53
SBERT+HC	**0.67**	**0.64**	**0.65**	**0.61**

Overlap: Proportion of ground truth points captured in generated summary

Precision: Proportion of generated points found in ground truth

3.3 Interactive Multi-Agent AI Framework

3.3.1 Simulating Multi-Perspective Business Meetings with AI

In our approach, we utilize the product summary from the previous section to simulate a business meeting environment. This simulation involves multiple LLM instances, each representing a different employee role within the organization. The roles and

their primary focus areas are as follows:

1. **Product Manager:** Prioritizes user engagement issues, defines project scope, coordinates interdepartmental efforts, aligns changes with business objectives, balances user needs with business goals, and establishes success metrics.

2. **UI/UX Designer:** Focuses on improving overall user experience, simplifying navigation, enhancing visual design, ensuring design consistency, incorporating accessibility features, and proposing innovative interaction patterns

3. **Technical Expert:** Addresses technical feasibility, performance impacts, codebase integration, scalability considerations, security implications, and potential technology stack updates

4. **Market Analyst:** Analyzes user behavior trends, evaluates competitive products, provides data-driven insights, identifies market opportunities and threats, forecasts user adoption rates, and suggests product differentiation strategies

Figure 3-9 shows the architectural diagram underlying our role-playing mechanism. We also attach an LLM at the end of the conversation to summarize and provide action items. We also created a web interface which users could directly interact with, as shown in 3-10. The user can view the pros and cons for each product - which are also passed in as context to the LLMs. They can also watch the LLMs converse with each other or directly ask questions to an agent.

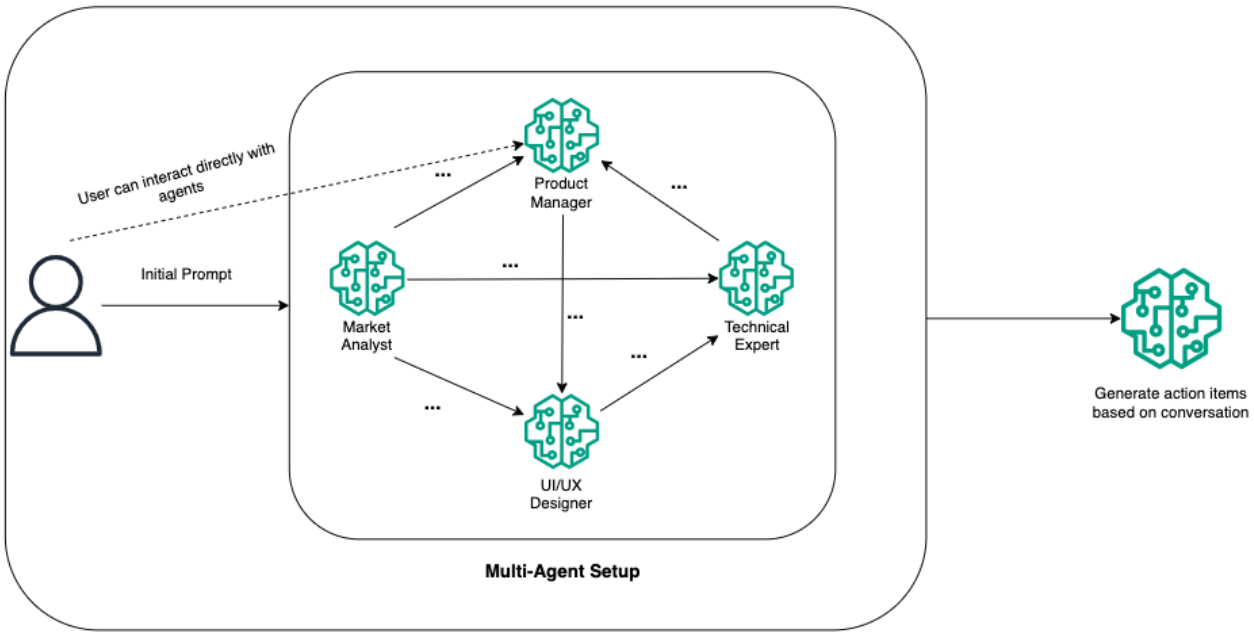

Figure 3-9: Architecture for role-playing mechanism

3.3.2 User-Driven Agent Interaction

Our framework extends beyond simulated multi-agent conversations by incorporating direct user interaction with individual AI agents. This feature allows users to engage in detailed dialogues with each role-specific agent, enabling a more nuanced exploration of ideas and critiques. Users can pose targeted questions to specific agents, such as asking the Technical Expert about implementation challenges, or request elaboration on particular points raised during the simulated meeting. Moreover, users have the flexibility to introduce new considerations or constraints for the agents to analyze, as well as challenge assumptions or propose alternative viewpoints.

This interactive capability transforms the tool from a passive ideation generator into an active collaborative partner, adapting to the user's specific needs and thought processes. It enables customized exploration, allowing users to guide the conversation towards areas of particular interest or concern, ensuring that the generated insights are directly relevant to their specific context. As users introduce new information or constraints, the AI agents can quickly adapt their recommendations, simulating

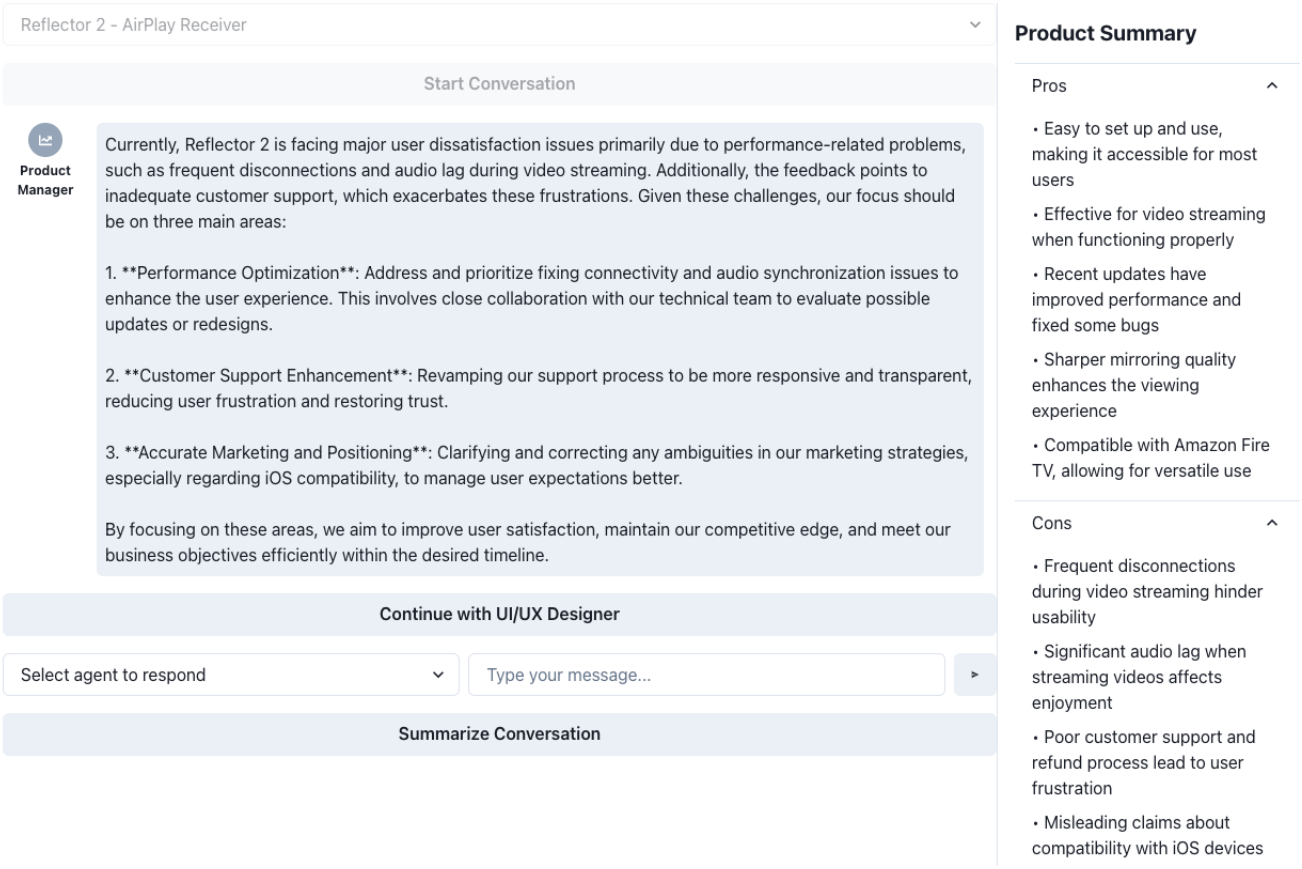

Figure 3-10: Web interface for users to interact with

real-time problem-solving scenarios.

The system provides both depth and breadth of analysis. Users can drill down into specific topics with individual agents while maintaining the broader multi-perspective view, allowing for both comprehensive and detailed analysis. This interaction also presents a continuous learning opportunity, as users gain insights into different professional perspectives, potentially broadening their own understanding and decision-making skills over time.

Furthermore, the framework supports flexible engagement, accommodating various levels of user involvement. Users can choose between passive observation of AI-generated discussions and active participation and direction, adapting to different work styles and needs. This flexibility extends to iterative refinement, where users can progressively refine ideas, testing different scenarios and adjusting parameters based on AI feedback and their own evolving understanding.

While the interactive multi-agent AI framework offers powerful capabilities, it is essential to view it as a complement to, rather than a replacement for, human expertise. The system serves as an advanced ideation and analysis tool, generating diverse perspectives and insights to inform human decision-making. Its true value lies in augmenting human creativity by providing novel connections and viewpoints, facilitating more comprehensive consideration of complex issues, and accelerating the initial stages of problem-solving and strategic planning.

Moreover, the rapid ideation capabilities of LLMs can significantly speed up the brainstorming and problem-solving processes within organizations. This efficiency not only saves valuable time but also enables the exploration of a much wider range of solutions, increasing the chances of finding truly innovative approaches to business challenges. By combining the analytical power and vast knowledge base of AI with human intuition, contextual understanding, and ethical judgment, organizations can leverage this framework to enhance their decision-making processes and drive innovation in an increasingly dynamic and competitive landscape.

3.3.3 Evaluation

The evaluation of conversational AI systems, particularly those involving multi-agent interactions, presents unique challenges due to the inherently qualitative and nuanced nature of the output. Traditional metrics often fail to capture the subtle complexities and contextual richness of business-oriented discussions. Therefore, we adopted an human-driven evaluation approach, leveraging Amazon Mechanical Turk (MTurk) to recruit workers who could provide comprehensive assessments of our system's performance across multiple dimensions.

3.3.3.1 Quantitative Analysis

To evaluate the quality of the business meeting simulation system, we recruited 20 participants through Amazon MTurk. Participants were required to have a task approval rate of >98% to ensure quality responses with attention checks incorporated. Each participant was asked to interact with three different products/simulated meetings and rate on various aspect of the system's performance.

Raters were provided with detailed rubrics for each evaluation metric. For example, when evaluating "Expert Role Simulation", a score of 1 indicated "The LLM responses appear generic and show no professional expertise," while a score of 5 indicated "The LLM consistently demonstrated deep domain knowledge and professional judgement comparable to industry experts." Similar detailed criteria were provided for each metric to ensure consistent evaluation.

Table 3.4: MTurk Worker Evaluation Scores (N=20, 3 meetings per worker)

Evaluation Metric	**Mean Score (1-5)**	**Std Dev**
Expert Role Simulation	4.47	0.62
Relevance to Discussion Points	4.40	0.71
Solution Feasibility	4.03	1.08
Coherence	4.02	1.09

The results demonstrate strong performance across all evaluated dimensions. The system scores particularly well in expert role simulation (M=4.47, SD=0.62), suggesting successful emulation of professional expertise. The relatively low standard

deviation indicates consistent performance across different scenarios. High scores in relevance (M=4.40, SD=0.71) and solution feasibility (M=4.03, SD=1.08) indicate that the system maintains focus on pertinent discussion points while generating practical solutions.

3.3.3.2 Qualitative Insights

The expert evaluation revealed several significant strengths of our system. A particularly notable finding was the system's comprehensive knowledge base, which enabled informed decision-making that closely mimicked real professional expertise. Experts repeatedly emphasized how the interactions felt authentically professional; closely resembling conversations with actual industry specialists.

Another strength was the system's interactive capability, allowing users to directly engage with the LLMs during the discussion. This feature provides enhanced control over conversation direction, marking a significant advancement over other similar frameworks.

The system's efficiency was another highlighted advantage, with experts noting that comprehensive meetings that traditionally require over an hour could be effectively simulated in 5-10 minutes, representing substantial time savings while maintaining discussion quality.

Experts also mentioned the system's ability to maintain focus on relevant topics and generate pertinent suggestions, demonstrating strong adherence to the primary discussion objectives.

Moreover, the intuitive user interface and clear workflow were consistently mentioned as factors contributing to the system's effectiveness, suggesting successful implementation of user-centric design principles.

3.3.3.3 Future Directions

While the current system demonstrates strong performance across multiple dimensions, several opportunities for enhancement exist. A primary area for future development lies in expanding the system's capability to handle more complex group

dynamics, particularly in scenarios involving conflict resolution and negotiation. This could be achieved through implementing more sophisticated turn-taking mechanisms and improving the system's ability to manage divergent viewpoints, ultimately creating more nuanced and realistic business discussions. For example, at some point in the conversation, there might be multiple thought processes you want to consider. One extension would be to create a tree of conversations, where you can discuss alternative paths in parallel.

Another improvement could be related to knowledge integration. Future iterations could incorporate real-time data integration capabilities, enabling the system to access more information about the company. For instance, the models could be given knowledge of the company's other products, planned improvements, etc.

There could also be improvements in the expertise profiles of the LLM agents. One focus could be on fine-tuning models to act like a certain professional. The difficulty with this approach is finding sufficient quality data to make this happen. Another area could be allowing the system to adapt to varied business environments and specialized domains.

Finally, the implementation of comprehensive documentation features presents a significant opportunity for enhancement. The current version has a primitive feature for generating summaries. Future versions could expand onto this and also include visualization tools for tracking discussion progress and decision-making patterns, making the system more valuable for post-meeting analysis and follow-up actions.

3.3.3.4 Additional Practical Applications of Multi-Agent Framework

Our work demonstrates a viable proof of concept for multi-agent LLM systems in business meeting simulations, showing that this approach can effectively generate diverse perspectives and facilitate complex discussions that traditionally require extensive cross-team collaboration. The success of our implementation suggests that similar frameworks could be valuable across numerous domains where multiple viewpoints and expertise are crucial for decision-making.

The potential applications of this framework span various sectors and use cases:

1. Banking fraud detection, involving fraud analysts, risk management specialists, compliance officers, and customer service representatives

2. Healthcare decision support, involving personas such as primary care physicians, specialists, pharmacists, and insurance coordinators

3. Urban planning initiatives, with architects, environmental scientists, traffic engineers, and community representatives

4. Educational curriculum development, incorporating perspectives from teachers, administrators, subject matter experts, and child development specialists

5. Supply chain optimization, featuring procurement specialists, logistics managers, quality control experts, and financial analysts

6. Product security assessments, including security engineers, penetration testers, privacy specialists, and user experience researchers

To demonstrate the framework's practical application and adaptability, we present a theoretical examination of its possible implementation in banking fraud detection, where the need for swift coordinated responses across multiple departments makes it an ideal use case for our approach. In the banking sector, fraud detection presents unique challenges that align well with our framework's capabilities. Traditional fraud detection processes often suffer from critical inefficiencies: communication delays between departments lead to slower response times, siloed decision-making results in incomplete risk assessment, and the lack of coordinated analysis can miss crucial patterns that only become apparent when viewing the situation from multiple angles [1].

Our multi-agent system addresses these challenges through a carefully designed set of specialized personas, each representing critical perspectives in fraud investigation. The fraud analyst persona focuses on pattern recognition and anomaly detection, bringing expertise in transaction analysis and emerging fraud techniques. The compliance officer ensures all responses align with regulatory requirements, including FinCEN guidelines and BSA requirements, while maintaining adherence to

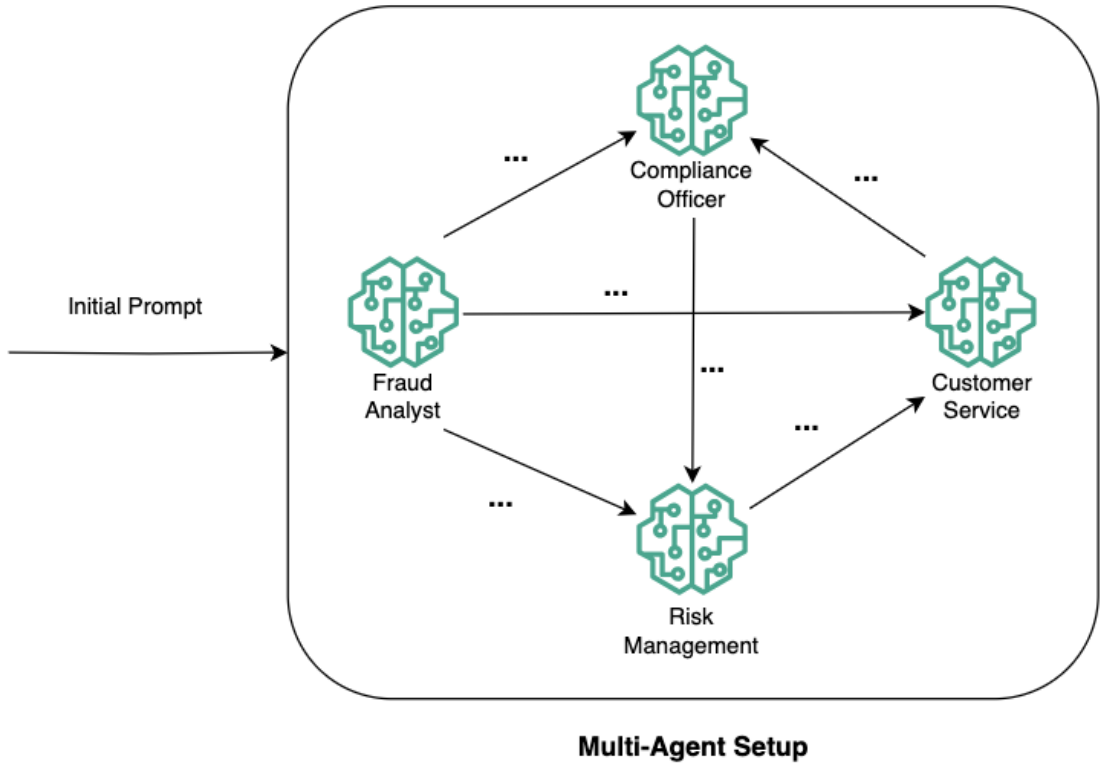

Figure 3-11: Our multi-agent framework applied to the problem of detecting fraud in a banking environment

internal policies [21]. The risk management specialist assesses potential financial exposure and proposes mitigation strategies, while the customer service representative provides crucial insights about customer behavior patterns and potential impact on client relationships.

The framework's modular architecture, as shown in Figure 3-11, enables a "plug and play" approach where organizations can easily customize the system by defining role-specific personas based on their unique requirements (note the similarity to Figure 3-9). The implementation maintains role-specific knowledge bases incorporating historical fraud patterns, regulatory requirements, customer interaction histories, and risk assessment matrices.

A key innovation is the system's parallel processing approach, where the fraud analyst's initial findings trigger immediate review by other personas. This enables simultaneous evaluation of regulatory compliance, customer impact, and financial risk

- achieving a level of efficiency impossible in traditional sequential workflows [12]. The system prioritizes rapid response while ensuring comprehensive analysis from multiple perspectives.

This framework could offer substantial organizational benefits that directly address key challenges faced by modern banking institutions. One of the most significant advantages is the reduction in coordination overhead. Traditional fraud response processes often require multiple meetings, email chains, and phone calls between different departments, creating delays and increasing the risk of miscommunication [1]. Our framework eliminates these inefficiencies by enabling immediate, parallel consultation between different organizational perspectives.

From a resource optimization standpoint, the framework allows banks to make better use of their expert staff. Instead of having fraud analysts spend time coordinating with other departments, they can focus on complex cases that require human expertise. The system effectively handles routine coordination and initial analysis, escalating cases to human experts only when necessary. This more efficient allocation of human resources can lead to significant cost savings while improving the overall quality of fraud detection and response [3].

The framework's ability to learn and adapt from each interaction creates a dynamic institutional knowledge base that captures cross-departmental expertise. This ensures consistent fraud detection capabilities even as staff changes occur and enables the system to evolve with emerging fraud patterns. Furthermore, the implementation can be enhanced with additional features to expand its capabilities. The system can incorporate real-time risk scoring that synthesizes insights from all personas, while automated regulatory compliance checking streamlines the handling of common scenarios. Advanced pattern recognition across historical cases helps identify emerging fraud trends, and customizable escalation thresholds allow organizations to align the system with their risk tolerance. The framework can also integrate seamlessly with existing fraud detection tools and databases, creating a comprehensive solution that builds upon established infrastructure.

Through this theoretical analysis, we explore how our proposed multi-agent LLM

framework could address specific industry challenges while maintaining its core advantages of parallel processing, comprehensive analysis, and efficient coordination. The approach suggests potential enhancements to operational efficiency while ensuring consistent consideration of diverse viewpoints, potentially leading to more robust and well-rounded solutions in fraud detection and beyond.

Chapter 4

A/B Testing

4.1 Introduction

In the realm of digital marketing and user experience optimization, A/B testing has long been considered the gold standard for hypothesis validation. However, as businesses strive for faster innovation and more efficient resource allocation, there is a growing need to re-evaluate our approach to hypothesis generation and testing.

A/B testing, while invaluable, can be a time consuming and resource-intensive process. According to recent statistics, 77% of companies are running A/B tests on their websites with 60% specifically testing their landing pages [6]. The investment in these tests is substantial, both in terms of time and financial resources. Microsoft, for instance, reports running over 1000 A/B tests on Bing search per month [6]. This level of testing, while thorough, can significantly delay the implementation of improvements and innovations.

The potential benefits of reducing A/B testing time are considerable. Bing improved its annual revenue per search by 10-25% due to A/B testing [6]. However, if we could achieve similar results in a fraction of time, the impact on business efficiency and profitability could be transformative. Moreover, with better UX design resulting from testing potentially increasing conversion rates by up to 400%, accelerating this process could lead to rapid and substantial improvements in user experience and business outcomes [6].

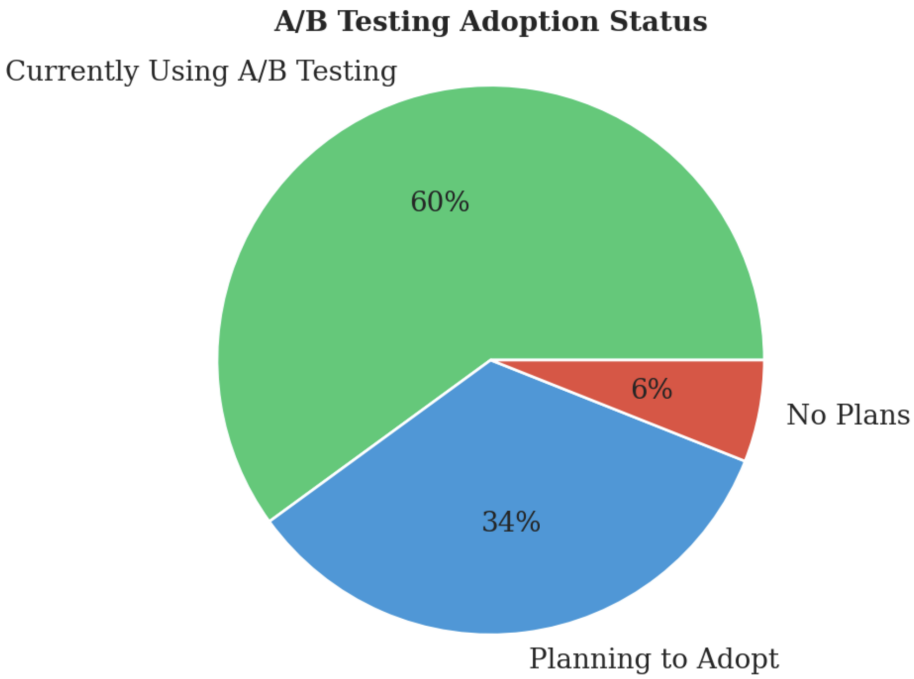

Figure 4-1: A/B Testing Adoption Status Among Companies [52]

Our approach to addressing this challenge is twofold. First, we develop a system that enables users to generate new headlines based on provided context, leveraging advanced NLP techniques. This will streamline the initial phase of headline creation, reducing the time spent on brainstorming and initial drafting.

Second, and more critically, we will focus on ranking these generated headlines to provide users with an immediate understanding of which options are likely to perform best. This ranking system will be based on a comprehensive analysis of linguistic features, psychological triggers, and historical performance data. By offering this predictive insight, we aim to significantly reduce the number of variants that need to be tested, thereby shortening the overall A/B testing cycle. This approach not only hopes to save time and resources but also to enhance the quality of headlines being tested.

4.2 Dataset

In this study, we leverage the Upworthy Research Archive, a comprehensive dataset that captures over 32,487 A/B tests conducted by Upworthy.com between January 2013 and April 2015[37]. Upworthy, a digital media company founded in 2012, pioneered viral content distribution with a focus on socially impactful stories, adopting the mission to bring attention to "stuff that matters" [37]. Their systematic approach to content optimization through rigorous A/B testing has inadvertently produced one of the largest publicly available datasets of headline performance metrics, providing valuable insights for research across disciplines including communications, political science, and psychology.

The dataset encompasses a rich collection of content optimization experiments, where each article underwent systematic testing through multiple variations of headlines, thumbnail images, and brief excerpts, collectively termed as "packages." Upworthy employed a robust randomized trial methodology, where website visitors were randomly assigned to different experimental conditions. The platform meticulously tracked engagement metrics, including clicks for each variation, total impressions, and temporal engagement patterns [37]. This systematic approach to data collection enables the analysis of "intent-to-treat" effects, providing insights into the causal impact of headline variations on user engagement, regardless of whether users actually viewed the content [37].

For the purposes of this study, we focus on a subset of key variables from the dataset: the test_id (unique identifier for each A/B test), headline variations, and the clicks and impressions for each headline. In order to develop an effective headline validation model, we first needed to address the fundamental challenge of quantifying relative headline performance. While our dataset contained raw clicks and impressions for each headline variant, we required a unified representation that could capture the statistical confidence of one headline's superiority over another.

4.2.1 Bayesian Click-Through Rate

For any pair of headlines (A, B) in our dataset, we observe:

$$\text{CTR}_A = \frac{\text{clicks}_A}{\text{impressions}_A} \tag{4.1}$$

$$\text{CTR}_B = \frac{\text{clicks}_B}{\text{impressions}_B} \tag{4.2}$$

However, these raw CTRs fail to account for the uncertainty inherent in the sampling process. To address this, we employed a Bayesian approach using Beta distributions, which serve as conjugate priors for binomial processes like click-through events. For each headline variant, we model the true CTR as a random variable following a Beta distribution:

$$\text{CTR}_{\text{true}} \sim \text{Beta}(\alpha, \beta) \tag{4.3}$$

where

$$\begin{aligned} \alpha &= \text{clicks} + 1 \\ \beta &= (\text{impressions} - \text{clicks}) + 1 \end{aligned} \tag{4.4}$$

The addition of 1 to both parameters represents Laplace smoothing, which provides regularization and prevents extreme probability estimates when sample sizes are small.

To determine the probability that one headline outperforms another, we employ Monte Carlo simulation. For headlines A and B, we generate N samples (typically $N \in \mathbb{N}, N = 100,000$) from their respective posterior distributions:

$$\begin{aligned} \text{samples}_A &\sim \text{Beta}(\text{clicks}_A + 1, \text{impressions}_A - \text{clicks}_A + 1) \\ \text{samples}_B &\sim \text{Beta}(\text{clicks}_B + 1, \text{impressions}_B - \text{clicks}_B + 1) \end{aligned} \tag{4.5}$$

The probability of headline A being superior to headline B is then estimated as:

$$P(A > B) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\text{samples}_A^{(i)} > \text{samples}_B^{(i)}] \tag{4.6}$$

where $\mathbf{1}[\cdot]$ is the indicator function.

This approach provides several advantages. It accounts for varying sample sizes between headline variants and provides natural uncertainty quantification. Also, it yields a single, interpretable metric $P(A > B) \in [0, 1] \subset \mathbb{R}$ that we can use for model training. This probability score serves as the foundation for our supervised learning approach (and a way to evaluate models). It provides a robust target variable that encapsulates both the magnitude and confidence of headline performance differences.

Moreover, the choice of pairwise comparisons over alternative approaches, such as multiclass classification or direct ranking, is motivated by several practical considerations. Since we experiment with both neural network based and LLM based approaches, we need a consistent methodology that works for both. A primary advantage of the pairwise approach is that it significantly increases our training data. For each A/B test with n variants, we can generate $\binom{n}{2}$ training examples, effectively amplifying our dataset size. While we could alternatively use a multiclass approach that simply labels the highest-performing headline, this would yield only one training example per test, substantially reducing the amount of learning signal available to the model.

Furthermore, this pairwise framework aligns naturally with the fundamental nature of A/B testing, where decisions are ultimately made through binary comparisons. By maintaining this connection to practical A/B testing methodology, our model's predictions remain directly applicable to real-world testing scenarios and more interpretable for practitioners. The pairwise approach also handles a common challenge in our dataset: variable test sizes. Different A/B tests may contain different number of variants, but any test with n variants can be consistently decomposed into pairs, providing a uniform framework for analysis regardless of the test size.

The pairwise approach also captures nuanced performance differences between headlines that might be lost in a multiclass framework. Rather than simply identifying

the best performer, our model learns to distinguish between relative performance between any two headlines, potentially capturing valuable insights about what makes one headline perform marginally better than another. This granular learning about relative performance difference could prove particularly valuable when generating and filtering new headline variants.

4.2.2 Exploratory Data Analysis

Our exploratory data analysis reveals several important insights about headline performance differences in A/B testing. Figure 4-2 presents two key visualizations that characterize the distribution of performances between headline variants. The density distribution (left panel) shows a distribution centered roughly at zero with a mean difference indicated by the red dashed line. This distribution suggests that while many headline pairs perform similarly, there exists substantial variation in relative performance. The right panel and 4-3 quantifies this variation, demonstrating that approximately 40% of all A/B tests in our dataset show significant differences, which we defined as probability differences greater than 0.1 from equal probability.

The detailed statistical measures presented in Table 4.1 provide additional context, with a median absolute difference of 0.079 and a mean absolute difference of 0.097. The standard deviation of 0.077 indicates consistent variability across tests, while the maximum observed difference of 0.500 highlights the potential for substantial improvements through optimal headline selection. These metrics underscore both the challenge and opportunity in headline optimization – while not all variants show meaningful differences, the significant proportion that do (40.1% showing significant differences) suggests that systematic approaches to headline generation and validation can yield meaningful improvements in content performance.

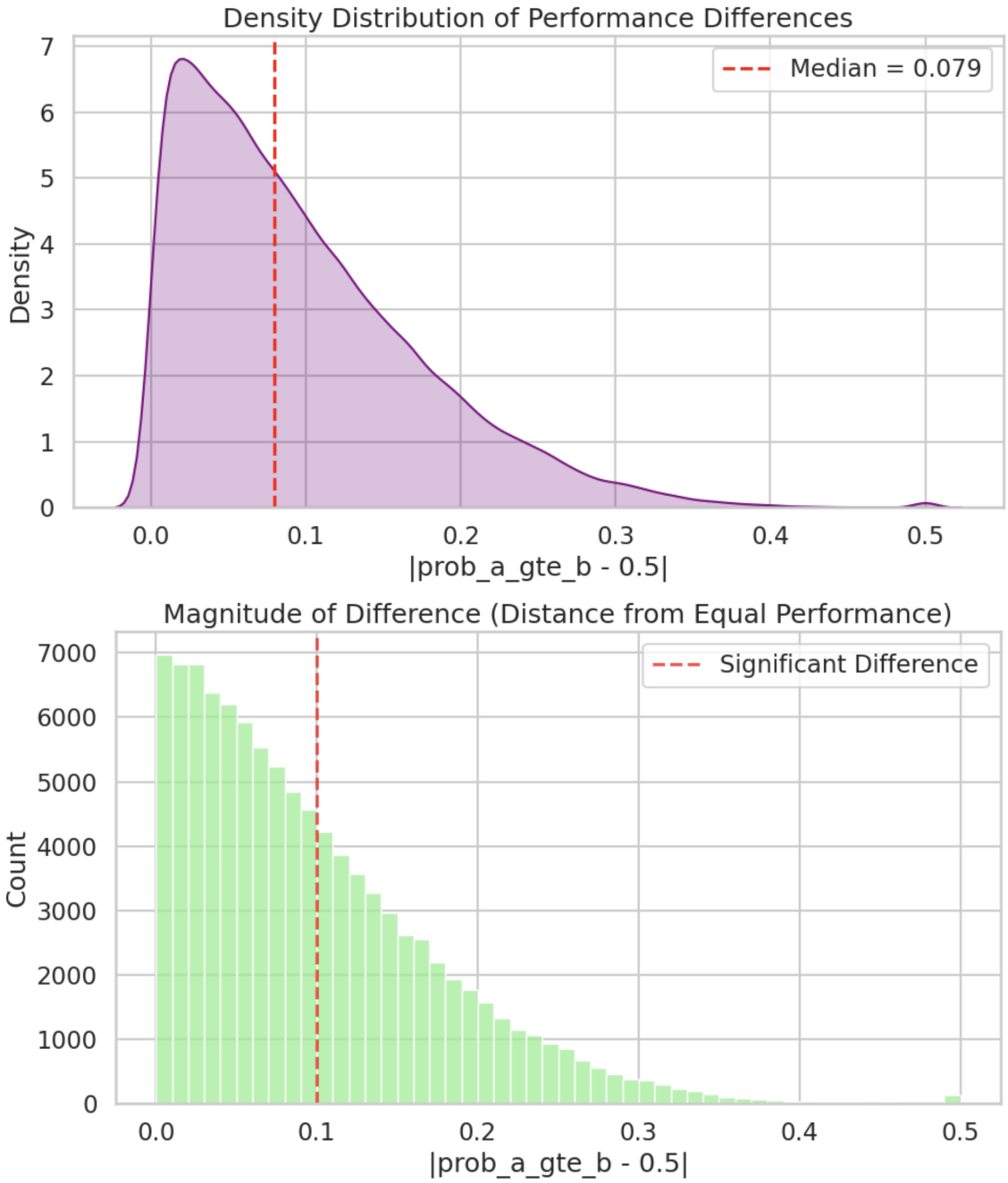

Figure 4-2: Analysis of probability differences in A/B tests. Top: Density distribution of performance differences with median marker (red dashed line). Bottom: Distribution of effect sizes, showing that approximately 40% of tests exhibit significant differences (>0.1 from equal probability).

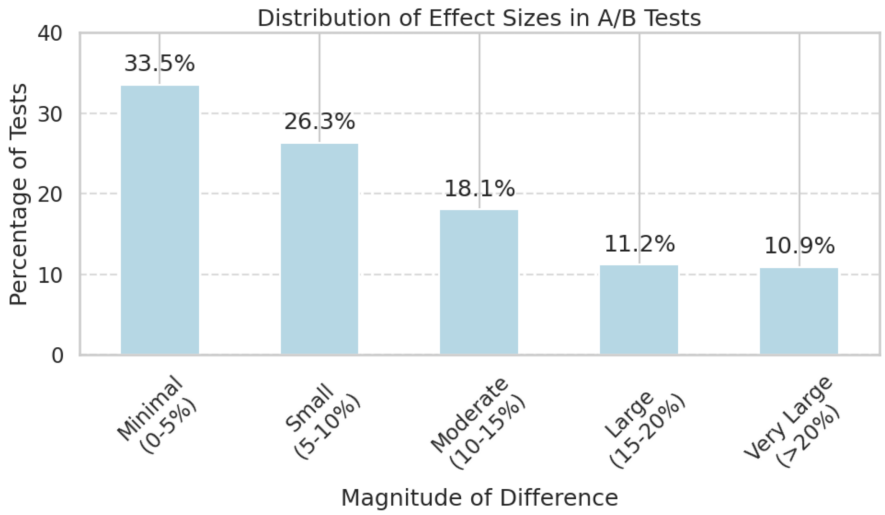

Figure 4-3: Large amount of experiments show at least a moderate (>10%) difference.

Table 4.1: A/B Test Statistical Analysis

Statistical Measure	**Value**
Median Absolute Difference	0.079
Mean Absolute Difference	0.097
Standard Deviation	0.077
Tests with Significant Difference	40.1%
Maximum Observed Difference	0.500

4.3 Methodology

In this section, we outline our approach to headline optimization through A/B testing. First, we describe a multi-LLM system for generating candidate headlines. Second, we present our headline validation framework, which considers both traditional machine learning and LLMs. Finally, we detail our evaluation metrics and experimental setup.

4.3.1 Generating Headlines for A/B Testing

Our headline generation system employs a multi-agent approach that leverages the complementary strengths of multiple LLMs working in concert. The system is design with flexibility and iterative improvement at its core, allowing for multiple rounds of generation and refinement to product high-quality candidate headlines.

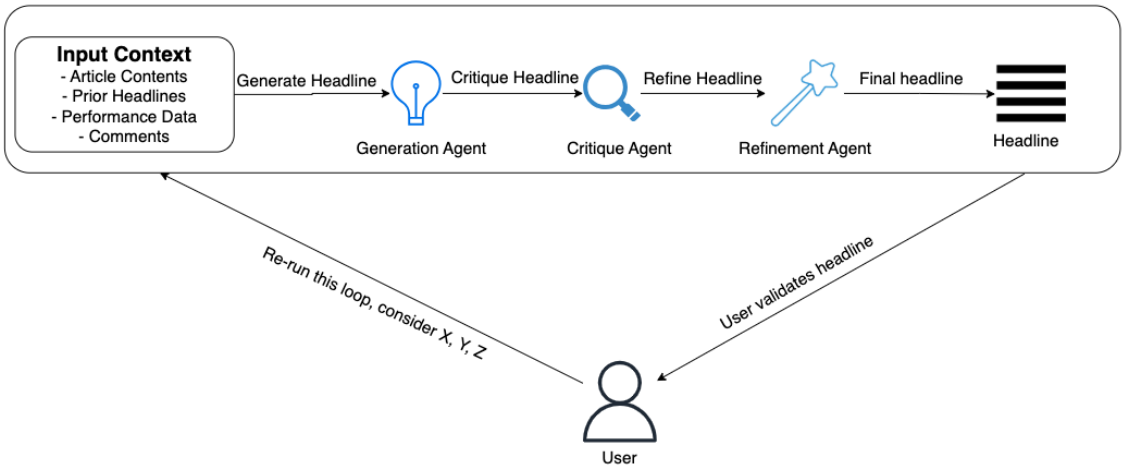

Figure 4-4: Multi-agent architecture to generate new headlines with user input

The process begins with a comprehensive input phase, where the system accepts not only the primary article context but also additional valuable information such as previously generated headlines, historical performance data, and specific content requirements. This flexible input structure allows the system to learn from past successes and adapt to evolving content strategies.

At the heart of our system is a three-agent architecture that orchestrates the headline generation process. The Generation Agent serves as the primary creative force, taking the input context and crafting initial headlines based on its understanding of effective headline writing principles and content relevance. This agent combines creativity with structured thinking to produce headlines that capture the essence of the content while maintaining engagement potential.

The Critique Agent then provides a sophisticated analysis of the generated headlines, examining multiple dimensions including clarity, engagement potential, accuracy, and brand voice consistency. Rather than simply identifying issues, this agent provides constructive feedback and specific suggestions for improvement. The critique phase is particularly valuable as it introduces an analytical perspective that helps refine and optimize the initial creative output.

The Refinement Agent represents the final stage of our primary generation cycle, synthesizing the original context, initial headlines, and critique feedback to produce

enhanced versions. This agent specializes in incorporating constructive feedback while maintaining the core message and appeal of the headlines.

A key advantage of our approach is its inherent flexibility. We support two options that allow for running multiple iterations. In the first option, the user can just configure a preset number of times to run. In the second option, we propose the final headline to the user and they can an option to re-run the system and add in additional comments. This iterative capability is particularly valuable when dealing with complex content or specific stylistic requirements. Additionally, the system can be configured to maintain a diverse pool of candidate headline, ensuring that different approaches and styles are represented in the final output.

However, we acknowledge a fundamental challenge in headline generation: the difficulty of predicting real-world performance without actual A/B testing. While our multi-agent system can generate headlines that adhere to best practices and incorporate expert knowledge, the true effectiveness of a headline can only be determined through user interaction and engagement metrics. This limitation leads us to our two-part solution: first, generate a diverse pool of candidate headlines using our flexible multi-agent system, and then apply our validation framework (detailed in the next section) to rank and select the most promising headlines for A/B testing.

4.3.2 Fine-tuned Models for Validation

4.3.2.1 DirectRank: A Direct Regression Approach to Headline Preference Learning

Our primary model architecture leverages the DeBERTa-v3 transformer [25] as its foundation, employing a siamese-style architecture to process and compare headline pairs. The model takes two headlines as input and outputs a probability indicating the likelihood that one headline will outperform the other in terms of user engagement.

The architecture can be formally expressed through a series of transformations. Given a pair of headlines H_A and H_B, the model first obtains contextual representations through the DeBERTa encoder:

$$\mathbf{h}_A = \text{DeBERTa}(H_A)[_{\text{CLS}}] \in \mathbb{R}^d \tag{4.7}$$

$$\mathbf{h}_B = \text{DeBERTa}(H_B)[_{\text{CLS}}] \in \mathbb{R}^d \tag{4.8}$$

where d represents the dimensionality of the DeBERTa hidden state. These representations are then concatenated and passed through a carefully designed regression head f_{reg} with multiple neural network layers:

$$P(H_A > H_B) = f_{\text{reg}}([\mathbf{h}_A; \mathbf{h}_B]) \tag{4.9}$$

The regression head employs a sequence of transformations with regularization:

$$\begin{aligned}
\mathbf{z}_1 &= \text{Dropout}(0.1)\big(\text{GELU}(\text{LayerNorm}(W_1[\mathbf{h}_A; \mathbf{h}_B] + b_1))\big) \\
\mathbf{z}_2 &= \text{Dropout}(0.1)\big(\text{GELU}(\text{LayerNorm}(W_2\mathbf{z}_1 + b_2))\big) \\
\hat{y} &= \sigma(W_3\mathbf{z}_2 + b_3)
\end{aligned}$$

where $W_1 \in \mathbb{R}^{256\times 2d}, W_2 \in \mathbb{R}^{64\times 256}, W_3 \in \mathbb{R}^{1\times 64}$ are learnable weight matrices, and σ represents the sigmoid activation function.

The model is trained using a composite loss function that combines mean squared error and mean absolute error:

$$\mathcal{L}(\hat{y}, y) = \text{MSE}(\hat{y}, y) + \lambda \text{MAE}(\hat{y}, y) \tag{4.10}$$

where $\lambda = 0.4$ balances the two loss components. To ensure numerical stability and proper convergence, we implement output clamping:

$$\hat{y}_{\text{final}} = \text{clamp}(\hat{y}, \epsilon, 1 - \epsilon) \tag{4.11}$$

where $\epsilon = 10^{-7}$. The model parameters are optimized using AdamW with weight decay and a linear learning rate schedule incorporating warmup. Training stability is

further enhanced through gradient clipping with a maximum norm of 1.0 and Xavier initialization with reduced gain for the regression layers.

This architecture learns to compare headlines by combining the contextual understanding of DeBERTa with a regularized regression mechanism, producing probability estimates for headline performance comparisons.

4.3.2.2 CAPS: Cross-Attention Preference Scorer

We propose another model that leverages DeBERTa-v3 architecture augmented with specialized cross-attention mechanisms [24] and explicit difference features for headline comparison tasks. Our model architecture introduces several key innovations in how it processes and compares pairs of headlines, moving beyond simple text encoding to capture nuanced relationships between headline variants.

The foundation of our model begins with encoding of headline pairs. For any two headlines (A, B), we first process their tokenized representations through a pretrained DeBERTa-v3-small model. This process is the same as 4.7 and 4.8.

To capture the intricate relationships between elements of both headlines, we implement a cross-attention mechanism [24]. This mechanism begins with the projection of our encoded representations into query, key, and value spaces:

$$\begin{aligned} Q_A &= H_A W_Q \\ K_B &= H_B W_K \\ V_B &= H_B W_V \end{aligned} \tag{4.12}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are learnable projection matrices. The cross-attention computation then proceeds as:

$$\text{Attention}(Q_A, K_B, V_B) = \text{softmax}\left(\frac{Q_A K_B^T}{\sqrt{d}}\right) V_B \tag{4.13}$$

This attention mechanism [54] is applied bidirectionally, computing both $\text{CrossAttn}(A \to B)$ and $\text{CrossAttn}(B \to A)$, allowing each headline to attend to the content of the other. The attention scores are normalized by $\sqrt{d}$ to stabilize gradients during train-

ing.

Next, our model incorporates explicit difference features through sequence comparison. For each headline pair, we compute a feature vector in $f \in \mathbb{R}^4$ that captures structural differences between the headlines:

$$f = \begin{bmatrix} \delta_{\text{changes}} \\ \mu_{\text{position}} \\ |\text{len}(A) - \text{len}(B)| \\ \rho_{\text{changes}} \end{bmatrix} \tag{4.14}$$

The components of this feature vector are computed as follows:

$$\begin{aligned} \delta_{\text{changes}} &= |\{\text{edit operations}\}| - 1 \\ \mu_{\text{position}} &= \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} p \\ \rho_{\text{changes}} &= \frac{\sum_i \max(|c_i^A|, |c_i^B|)}{\max(|A|, |B|)} \end{aligned} \tag{4.15}$$

where $\mathcal{P}$ represents the set of positions where changes occur, c_i^A and c_i^B represent the changed segments in headlines A and B respectively.

The final stage of our model integrates all extracted features through a series of non-linear transformations. We first concatenate all our representations:

$$h = [h_A; h_B; c_{AB}; c_{BA}; E(f)] \tag{4.16}$$

where h_A and h_B are the pooled representations of each headline, c_{AB} and c_{BA} are the pooled cross-attention outputs, and $E(f)$ represents the encoded difference features. The final classification proceeds through multiple layers:

$$\begin{aligned} z_1 &= \text{LN}(W_1 h + b_1) \\ z_2 &= \text{ReLU}(z_1) \\ z_3 &= \text{Dropout}(z_2, p = 0.1) \\ p &= \sigma(W_2 z_3 + b_2) \end{aligned} \tag{4.17}$$

where LN denotes layer normalization, σ is the sigmoid activation function, and $p \in [0, 1]$ represents the final probability score indicating the relative performance of headline A compared to headline B.

4.3.3 LLM-based Approaches for Validation

This sections presents our methodological approach to predicting headline engagement using LLM techniques. Our research explores increasingly sophisticated prompting strategies and model combinations to determine the most effective approach for predicting relative headline performance.

4.3.3.1 Zero-Shot Prediction Approaches

Our initial methodology employs zero-shot prediction, both with and without chain-of-thought (COT) reasoning. This approach tests the model's inherent ability to understand headline effectiveness without example training data. The COT variant encourages explicit reasoning about various headline characteristics.

Prompt 4: Zero-Shot Base Prediction

```
System: You are an expert headline analyst with deep expertise in
digital media, user psychology, and content optimization. Your role is
to evaluate headline effectiveness based on established principles of
user engagement and click-through rate optimization.
User: Analyze these two headlines for their potential reader
engagement and click-through rate
Headline A: {headline_a}
Headline B: {headline_b}
Consider factors including but not limited to:

  1. Emotional resonance and psychological triggers

  2. Information clarity and value proposition

  3. Curiosity gap optimization
```

```
  4. Use of power words and engagement patterns

  5. Length and readability optimization

Based on these factors, output only 'A' or 'B' to indicate which
headline would perform better.
```

Prompt 5: Zero-Shot Chain-of-Thought

```
<Same base prompt as 4>
Analyze each factor step by step, then provide your final answer in
this JSON format:

{
    "analysis": {
        "emotional_impact": "analysis of emotional elements",
        "clarity": "analysis of information structure",
        "curiosity": "analysis of curiosity triggers",
        "language": "analysis of word choice and power words",
        "optimization": "analysis of technical factors"
    },
    "comparative_reasoning": "step-by-step comparison",
    "prediction": "A or B"
}
```

4.3.3.2 Few-Shot Learning Approaches

Building upon the zero-shot framework, we implemented few-shot learning techniques with various enhancements including COT reasoning and retrieval augmented generation (RAG) [23]. This methodology provides the model with carefully selected examples to improve prediction accuracy.

Prompt 6: Few-Shot Prediction

```
<Same base prompt as 4>
Study these example headline pairs and their performance metrics:
{examples_formatted_with_metrics}
```

Prompt 7: Few-Shot with COT

```
<Same base prompt as 5>
Study these example headline pairs and their performance metrics:
{examples_formatted_with_metrics}
```

For the RAG based approach, we use the SBERT model, all-MiniLM-L6-v2, to generate dense vector representations of headline pairs. This approach captures semantic relationships between headlines beyond simple lexical matching, allowing for more nuanced similarity comparisons. Each headline pair is encoded as a combined embedding, preserving the contextual relationship between headlines A and B.

The system computes cosine similarity between the test headline pair and all training examples. This comparison occurs in the high-dimensional embedding space, allowing for sophisticated matching that captures subtle semantic relationships. We provide the top 5 closest results as part of the prompt.

Prompt 8: Few-shot with COT + RAG

```
<Same base prompt as 5>
Study these example headline pairs and their performance metrics:
{top_5_closest_examples}
```

4.3.3.3 Multi-Agent Approach

We also experimented with another approach that employs multiple specialized prompts targeting different aspects of headline effectiveness, implemented across multiple LLM models for comprehensive analysis. We use a linguistic agent that focuses on headline optimization and the effect of words on user engagement. Concurrently, we also ask a psychological agent that is an expert in user engagement and content consumption

behavior. Lastly, we take the output of both agents and pass it into a synthesizer agent, which also considers the RAG examples.

Prompt 9: Linguistic Expert

```
System: You are a linguistic expert specializing in headline
optimization. Focus exclusively on technical and structural elements
of headline composition.
User: Analyze these headlines focusing on structural elements:
Headline A: {headline_a}
Headline B: {headline_b}
Provide analysis in this JSON format:

{
    "structural_analysis": {
        "headline_a": {
            "syntax_score": <1-10>,
            "clarity_score": <1-10>,
            "information_hierarchy": "analysis of structure",
            "technical_elements": ["list", "of", "elements"],
            "structural_strengths": ["list", "of", "strengths"],
            "structural_weaknesses": ["list", "of", "weaknesses"]
        },
        "headline_b": { ... },
    "comparative_analysis": "structural comparison",
    "structural_winner": "A or B",
    "confidence": <float between 0 and 1>
    }
}
```

Prompt 10: Psychological Analysis

```
System: You are a psychological expert in user engagement and content
consumption behavior.
```

```
User:  Analyze these headlines focusing on psychological impact:
Headline A: {headline_a}
Headline B: {headline_b}
Provide analysis in this JSON format:

{
    "psychological_analysis": {
        "headline_a": {
            "emotional_impact": <1-10>,
            "curiosity_generation": <1-10>,
            "psychological_triggers": ["list", "of", "triggers"],
            "engagement_mechanics": ["list", "of", "mechanics"],
            "behavioral_predictions": ["list", "of", "predictions"]
        },
        "headline_b": { ... },
        "comparative_analysis": "psychological comparison",
        "psychological_winner": "A or B",
        "confidence": <float between 0 and 1>
    }
}
```

Prompt 11: Final Synthesis

```
System:  You are a senior editor synthesizing multiple expert analyses
to make final headline performance predictions.
User:  Consider these expert analyses for the following headlines:
Headline A: {headline_a}
Headline B: {headline_b}
Structural Analysis:  {structural_analysis_json}
Psychological Analysis:  {psychological_analysis_json}
Previous similar examples:  {top_5_similar_examples}

Provide your synthesis in this JSON format:
```

```
{
   "synthesis": {
       "headline_a": {
           "combined_score": <1-10>,
           "key_advantages": ["list", "of", "advantages"],
           "key_disadvantages": ["list", "of", "disadvantages"]
       },
       "headline_b": { ... },
       "cross_analysis_patterns": ["list", "of", "patterns"],
       "historical_alignment": "analysis of alignment with examples"
   },
   "final_prediction": {
       "winner": "A or B",
       "confidence": <float between 0 and 1>,
       "reasoning": "detailed synthesis explanation"
   }
}
```

Each prompt system was implemented using both GPT-4 and Claude models, allowing for comparison and validation across different LLM architectures.

4.4 Evaluation

Our experimental results reveal significant variations in performance across different methodologies for headline effectiveness prediction. The DirectRank baseline method achieves an accuracy of 0.54, performing only marginally better than random guessing. Analysis of DirectRank's probability distributions shows predictions clustering around 0.5, indicating low confidence and poor discriminative ability. In contrast, the CAPS method demonstrates strong performance with an accuracy of 0.81, validating the effectiveness of our cross-attention mechanism and feature selection approach in capturing relevant patters for headline comparison.

Table 4.2: Headline Prediction Accuracies by Method and Model

Method	**GPT-4**	**Claude**	**Other**
Baseline Methods			
DirectRank	–	–	0.54
CAPS	–	–	0.81
Zero-Shot Methods			
Basic	0.62	0.62	–
Chain-of-Thought	0.62	0.70	–
Few-Shot Methods			
Basic	0.60	0.62	–
Chain-of-Thought	0.58	0.70	–
COT + RAG	0.80	**0.86**	–
Advanced Methods			
Multi-LLM	0.74	0.82	–

The zero-shot and basic few-shot approaches both with and without COT reasoning, show moderate performance with accuracies ranging from 0.58 to 0.70. This middling performance suggests that while LLMs possess some inherent understanding of headline effectiveness, they struggle to make consistent predictions without proper contextualization or relevant examples. The similar performance across these methods indicates that simply adding reasoning steps or a few examples without careful selection does not significantly enhance the models' predictive capabilities.

The best performance comes from the few-shot COT + RAG approach, particularly with Claude, achieving an impressive 0.86 accuracy. This superior performance demonstrates the crucial importance of providing LLMs with semantically relevant examples. The RAG methodology's success can be attributed to its ability to dynamically select and present the most pertinent historical examples, allowing the model to identify and apply patterns from truly similar cases rather than relying on generic headline writing principles.

The multi-LLM approach also shows strong performance, particularly with Claude (0.82), indicating that combining multiple specialized perspectives (structural, psychological, and synthesis) provides robust prediction capabilities. This suggests that breaking down the analysis into distinct aspects and then synthesizing leads to more reliable predictions than single-perspective approaches. Moreover, this approach is

highly interpretable and gives the user insight into the decision-making process of the LLMs.

Notably, Claude consistently outperforms GPT-4 across most experimental conditions, with particularly pronounced differences in the more sophisticated approaches (Few-shot RAG: 0.86 vs 0.80, Multi-LLM: 0.82 vs 0.74). This performance gap suggests that Claude may have better understanding of language patterns and user engagement factors, or that its architecture is better suited for comparative analysis tasks.

While our system provides pairwise headline comparisons to assist content creators, we acknowledge that these binary predictions may not always yield a globally consistent ordering (i.e., if headline A is predicted to outperform B, and B outperforms C, the system might still predict C to outperform A). Our current implementation defers to human judgment in such cases, aligning with our goal of augmenting rather than replacing human decision-making. Future work could address this limitation in several ways: (1) extending the model to output confidence scores for each prediction, then applying algorithms like Bradley-Terry to derive a complete, probabilistic ranking that minimizes inconsistencies across all pairwise comparisons; (2) expanding our multi-agent workflows where specialized LLM agents evaluate different aspects of headlines (e.g., emotional appeal, informativeness, clarity) and collaborate through structured communication protocols to reach consensus on multi-criteria rankings, potentially using techniques from preference aggregation theory to combine their individual assessments into a coherent global ordering.

In conclusion, our current system and approach is a good start to reducing the amount of resources required for A/B testing; however, additional work could be done to further enhance our system.

Chapter 5

Conclusion

This thesis has explored two significant applications of LLMs in product development, demonstrating their potential to transform traditional processes and methodologies. Our first project introduced a comprehensive approach to enhancing product innovation discussions through a two-phase methodology. The initial phase tackled the challenge of distilling insights from large-scale customer feedback, combining SBERT embeddings with hierarchical clustering and LLM-powered summarization. When tested on the Amazon reviews dataset, our SBERT+HC approach demonstrated exceptional performance across both automated metrics and human evaluation, achieving BERTScores of 0.89 (precision), 0.88 (recall), and 0.86 (F1). Human evaluation further validated these results, with our method achieving 67% overlap for pros and 65% for cons with ground truth summaries. Particularly noteworthy was the system's ability to identify and surface nuanced product features that are often overlooked in traditional summary approaches, demonstrating more thorough analysis than human summarizers and providing product teams with richer, more comprehensive insights for decision-making.

Building upon these condensed insights, we developed a multi-agent LLM framework that successfully simulated professional product development discussions. The strong validation through Amazon Mechanical Turk evaluations, with scores of 4.47/5 for expert role simulation, 4.40/5 for discussion relevance, and 4.03/5 for solution feasibility, suggests that this approach could improve how product teams collaborate and

iterate on ideas. Qualitative feedback particularly highlighted the system's ability to authentically simulate professional expertise while significantly reducing the time required for comprehensive discussions from hours to minutes. The framework's ability to maintain consistent professional perspectives while enabling meaningful cross-functional dialogue opens new possibilities for rapid prototyping of product ideas and more efficient decision-making processes. This framework demonstrated remarkable versatility, as evidenced by its successful adaptation to complex business scenarios beyond product development.

A particularly compelling exploration of our framework's adaptability was our detailed case study of its potential application to banking fraud detection. We outlined how the multi-agent system could significantly enhance existing processes by simulating specialized roles including fraud analysts, compliance officers, and risk management specialists, enabling simultaneous multi-perspective analysis of potential fraud cases. Our analysis identified several potential benefits: reduction in response times through streamlined cross-departmental communication, more comprehensive risk assessment through integrated expert perspectives, and improved resource allocation by automating routine coordination while preserving human oversight for complex cases. This theoretical case study demonstrates the framework's potential for addressing real-world challenges across different domains, particularly in scenarios requiring rapid, coordinated responses from multiple expert perspectives.

The second major contribution of this thesis focused on optimizing A/B testing processes through a two-stage methodology. Traditional A/B testing, while effective, often requires significant time and resources to validate content decisions. Our approach leverages LLMs: first generating diverse headline candidates, and then employing a validation methodology to predict performance without the need for extensive live testing. Our few-shot Chain-of-Thought with RAG approach achieved 86% accuracy in predicting relative headline performance on the Upworthy dataset, significantly outperforming baseline methods and simpler approaches. The multi-LLM approach also showed strong results with 82% accuracy, demonstrating the value of combining multiple specialized perspectives in content evaluation. These results

demonstrate the potential for significantly reducing dependency on traditional A/B testing methods while maintaining high-quality decision-making.

The implications of this A/B testing optimization extend beyond mere efficiency gains. By enabling rapid, accurate predictions of content performance, companies can iterate more quickly on their content strategy while maintaining high confidence in their decisions. This capability is particularly valuable in today's fast-paced digital environment, where the ability to quickly test and validate content can provide a significant competitive advantage. The methodology we developed could potentially save organizations substantial resources by reducing the need for lengthy A/B tests while still providing reliable guidance for content optimization decisions.

Looking broadly at both projects, this research demonstrates the transformative potential of LLMs in business processes when applied thoughtfully and systematically. Our work suggests that LLMs can do more than simply generate content or respond to queries—they can serve as sophisticated tools for analysis, simulation, and prediction. The real-world impact of these contributions is tangible:

1. For product development teams, our review analysis and multi-agent discussion system provides a more efficient way to process customer feedback and simulate cross-functional collaboration, potentially reducing the time from insight to action.

2. For financial institutions, our fraud detection implementation demonstrates how multi-agent systems can enhance existing processes, improving response times and decision quality while optimizing resource utilization.

3. For content creators and marketers, our A/B testing optimization approach offers a more efficient path to content validation, potentially reducing time-to-market while maintaining decision confidence.

Future work could explore several promising directions. The multi-agent framework could be enhanced with more specialized roles and domain-specific knowledge, while the integration of real-time data could further improve its adaptability. For

the A/B testing system, extending the methodology to other types of content beyond headlines, such as product descriptions or user interface elements, could provide broader value. Additionally, investigating the integration of these different approaches could lead to a unified platform that combines customer insight analysis, cross-functional collaboration, and predictive testing into a seamless workflow.

These contributions open new avenues for research at the intersection of artificial intelligence and business processes. While our results are promising, they also raise important questions about the evolution of organizational practices in an AI-enhanced future. How might these tools change the nature of team collaboration? What new skills and processes will professionals need to develop to effectively leverage these capabilities? How can organizations balance the efficiency gains of AI automation with the need for human judgment and oversight? As LLM technology continues to advance, addressing these questions while building upon the methodologies presented in this thesis will be crucial for realizing the full potential of AI in transforming business operations.

The practical implementations and positive results demonstrated in this thesis suggest that we are at the beginning of a significant transformation in how organizations can leverage AI to enhance their operations. By providing concrete methodologies and frameworks that bridge the gap between AI capabilities and business needs, this work contributes to the foundation of what may become standard practices in AI-enhanced business operations.

Bibliography

[1] Hanae Abbassi, Imane El Alaoui, and Youssef Gahi. Fraud detection techniques in the big data era. *Scitepress*, 2021.

[2] Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[3] Thushara Amarasinghe, Achala Aponso, and Naomi Krishnarajah. Critical analysis of machine learning based approaches for fraud detection in financial transactions. In *Proceedings of the 2018 International Conference on Machine Learning Technologies*, ICMLT '18, pages 12–17, New York, NY, USA, 2018.

[4] Guy Azov, Tatiana Pelc, Adi Fledel Alon, and Gila Kamhi. Self-improving customer review response generation based on LLMs. *arXiv preprint arXiv:2405.03845*, May 2024.

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, 2021.

[6] Ivan Blagojević. A/b testing statistics. 99 firms, https://99firms.com, 2024.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[8] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Sanjeev Arora, Sydney von Arx, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[9] James Brand, Ayelet Israeli, and Donald Ngwe. Using llms for market research. Technical Report 23-062, Harvard Business School Working Paper, July 2024.

[10] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[11] Alexander Changeux and Stephen Montagnier. Strategic decision-making support using large language models (llms). *Management Journal for Advanced Research*, 4(4):102–108, August 2024.

[12] Z. Chen, S. Zhang, X. Zeng, M. Mei, X. Luo, and L. Zheng. Parallel path detection for fraudulent accounts in banks based on graph analysis. *PeerJ Comput. Sci.*, 9:e1749, 2023.

[13] Srivas Chennu, Andrew Maher, Christian Pangerl, Subash Prabanantham, Jae Hyeon Bae, Jamie Martin, and Bud Goswami. Rapid and scalable bayesian ab testing. *arXiv preprint arXiv:2307.14628*, 2023.

[14] Ming Cheung. A reality check of the benefits of llm in business. *arXiv preprint arXiv:2406.10249*, 2024.

[15] Aakanksha Chowdhery et al. PaLM: Scaling language modeling with pathways. *arXiv preprint*, 2022.

[16] Statista Research Department. Selected services reaching 100 million followers as of August 2024. *Statista*, 2024.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, May 2023.

[19] Eva Eigner and Thorsten Händler. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*, 2024.

[20] Taha Falatouri, Denisa Hrušecká, and Thomas Fischer. Harnessing the power of llms for service quality assessment from user-generated content. *IEEE Access*, 12:99755–99767, 2024.

[21] Financial Crimes Enforcement Network (FinCEN). Fincen overview. https://www.fincen.gov/what-we-do, 2024.

[22] Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*, March 2023.

[23] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2023.

[24] Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pretrained transformers for machine translation. In Marie-Francine

Moens, Xuanjing Huang, Lucia Specia, and Scott Wen tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765, Online and Punta Cana, Dominican Republic, Nov 2021.

[25] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[26] David G. Jansson and Steven M. Smith. Design fixation. *Design Studies*, 12(1):3–11, 1991.

[27] Olivier Jeunen and Aleksei Ustimenko. Learning metrics that maximise power for accelerated a/b-tests. *arXiv preprint arXiv:2402.03915*, 2024.

[28] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5093, 2020.

[29] Haolin Jin, Linghan Huang, Haipeng Cai, Jun Yan, Bo Li, and Huaming Chen. From llms to llm-based agents for software engineering: A survey of current, challenges and future. 2024.

[30] Jana Juroš, Laura Majer, and Jan Šnajder. Llms for targeted sentiment in news headlines: Exploring the descriptive-prescriptive dilemma. *arXiv preprint arXiv:2403.00418*, 2024.

[31] Jeffrey Kuiken, Anne Schuth, Martijn Spitters, and Maarten Marx. Effective headlines of newspaper articles in a digital environment. *Digital Journalism*, 5(10):1300–1314, 2017.

[32] N. Larsen, J. Stallrich, S. Sengupta, A. Deng, R. Kohavi, and N. T. Stevens. Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician*, 78(2):135–149, 2023.

[33] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004.

[34] Bing Liu. *Sentiment analysis and opinion mining*, volume 5 of *Synthesis lectures on human language technologies*. Morgan & Claypool Publishers, 2012.

[35] Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-Yi Lee, and Shao-Hua Sun. LLM discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*, May 2024.

[36] T. J. Marion, M. Srour, and F. Piller. When generative AI meets product development. *MIT Sloan Management Review*, 66(1):14–15, 2024.

[37] J.N. Matias, K. Munger, M.A. Le Quere, et al. The upworthy research archive, a time series of 32,487 experiments in u.s. media. *Scientific Data*, 8:195, 2021.

[38] Lukas Meincke, Karan Girotra, Gideon Nave, Christian Terwiesch, and Karl T. Ulrich. Using large language models for idea generation in innovation. *SSRN*, 2023.

[39] OpenAI, Josh Achiam, Steven Adler, et al. GPT-4 Technical Report. *arXiv preprint*, 2023.

[40] Matteo Ottaviani, Stefan M. Herzog, Pietro Leonardo Nickl, and Philipp Lorenz-Spreen. How a/b testing changes the dynamics of information spreading on a social network. *arXiv preprint arXiv:2405.01165*, 2024.

[41] Martin Potthast, Tim Gollub, Matthias Hagen, and Benno Stein. The clickbait challenge 2017: Towards a regression model for clickbait strength. *arXiv preprint arXiv:1812.10847*, 2018.

[42] Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation. *arXiv preprint arXiv:2407.08940*, July 2024.

[43] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[44] Sashank J. Reddi and Sanjiv Kumar. Efficient training of language models using few-shot learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 14553–145682. PMLR, 2023.

[45] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 2019.

[46] Md Main Uddin Rony, Md Mahfuzul Haque, Mohammad Ali, Ahmed Shatil Alam, and Naeemul Hassan. Exploring the potential of the large language models (llms) in identifying misleading news headlines. In *Proceedings of the 1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems*, pages 1–5, Honolulu, HI, USA, May 2024.

[47] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056, 2024.

[48] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org, 2023.

[49] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*, October 2023.

[50] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train Megatron-Turing NLG 530B, a large-scale generative language model. *arXiv preprint*, 2022.

[51] Yun-Zhu Song, Hong-Han Shuai, Sung-Lin Yeh, Yi-Lun Wu, Lun-Wei Ku, and Wen-Chih Peng. Attractive or faithful? popularity-reinforced learning for inspired headline generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2020)*, Taiwan, 2020. Association for the Advancement of Artificial Intelligence (AAAI). AAAI 2020.

[52] Taras Talimonchuk. What is a/b testing: Steps & best practices. Claspo Blog, https://claspo.io, 2024.

[53] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*, June 2024.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[55] Chao Wang, Xiaoyan Jiang, Qing Li, Zijuan Hu, and Jie Lin. Leveraging LLM Agents to Extract Customer Needs from User-Generated Content, October 2024.

[56] Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*, September 2023.

[57] Jie JW Wu. Autooffab: Automated offline a/b testing for requirement engineering. In *Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, FSE 2024, pages 472–476, New York, NY, USA, 2024.

[58] Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. Clickbait? sensational headline generation with auto-tuned reinforcement learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[59] Zikun Ye, Hema Yoganarasimhan, and Yufeng Zheng. Lola: Llm-assisted online learning algorithm for content experiments. *arXiv preprint arXiv:2406.02611*, 2024.

[60] Jianyi Zhang, Xu Ji, Zhangchi Zhao, Xiali Hei, and Kim-Kwang Raymond Choo. Ethical Considerations and Policy Implications for Large Language Models: Guiding Responsible Development and Deployment, 2023.

[61] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.

[62] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[63] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[64] Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*, April 2024.

[65] Yuxiang Zhu, Joao R. A. Moniz, Siddharth Bhargava, Junda Lu, Divya Piraviperumal, Shixiang Li, Yuhao Zhang, Hong Yu, and Bo Tseng. Can large language models understand context? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018. Association for Computational Linguistics, 2024.

[66] Arkaitz Zubiaga. Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence*, 6:1350306, January 2024.