Yönetim Bilimleri Dergisi/Journal of Administrative Sciences
Cilt/Volume: 23, Sayı/No: 57, ss./pp.: 1705-1725
DOI: https://doi.org/10.35408/comuybd.1595049

-RESEARCH ARTICLE-

HOW TO USE GENAI AS A RATER IN MARKETING: A COMPREHENSIVE GENAI-BASED CONTENT CODING FRAMEWORK

Bahri Baran KOÇAK[1]

Abstract

With the recent developments in Generative AI (GenAI) applications, popular tools such as ChatGPT have gained potential not only for individual or commercial purposes but also for the marketing discipline. As an important process for the qualitative data analysis in marketing research, the task of coding texts is mostly performed by human experts, and this can cause loss of time and increase costs. These tools may also have the potential to perform the task of coding texts, which is mostly performed by human experts for the qualitative data analysis process in marketing research. Although there are techniques applied for coding qualitative data, no study has yet presented a systematic framework to use GenAI tools in coding tasks, especially for marketing research. To fill this research gap, the current study proposes a 6-step GenAI-based content coding framework. In the framework, firstly brand messages are collected, and the themes needed for classification are determined. Then, appropriate Gen-AI models are selected to separate brand messages according to the desired themes, prompts are prepared, classes with AI outputs are coded and finally, compatibility between coders is checked. In this respect, an application is carried out to test the proposed framework. Informational, entertaining and remunerative content strategies in consumer engagement literature were used as themes and the coding agreements between 3 Large Language Models (LLMs) and human experts were determined with Kappa statistics. According to the results, the level of agreement between Human and ChatGPT gave the best Inter-Rater Reliability (IRR) for informational and remunerative contents in the comparison of Human and AI coders, while Gemini performed better for entertaining messages. As a most important practical contributions, the framework of this study offers a useful and faster coding processes for marketing practitioners.

Keywords: *GenAI, large language model, inter-rater reliability, marketing, framework.*

JEL Codes: *M31, D83.*

Başvuru: *02.12.2024* **Kabul:** *04.04.2025*

[1] Dr., Dicle Üniversitesi, Sivil Havacılık Yüksek Okulu, Havacılık Yönetimi Anabilim Dalı, Diyarbakır, Türkiye, bahribaran.kocak@dicle.edu.tr, ORCID: 0000-0001-5658-7371

PAZARLAMADA YAPAY ZEKAYI DEĞERLENDİRİCİ OLARAK NASIL KULLANIRIM? ÜRETKEN AI İLE KAPSAMLI BİR İÇERİK KODLAMA ÇERÇEVESİ[2]

Öz

Yapay zekâ (AI) alanındaki son gelişmelerle birlikte, üretken-AI (GenAI) uygulamaları arasında öne çıkan ChatGPT gibi popüler araçlar, bireysel veya ticari amaçların yanı sıra pazarlama disiplini için de potansiyel vadetmektedir. Pazarlama araştırmasında metin kodlama görevi, nitel veri analizi için önemli bir süreç olmakla birlikte çoğunlukla insanlar tarafından gerçekleştirilmektedir ve bu durum vakit kaybına ve maliyet artışına neden olabilmektedir. GenAI araçları, pazarlama araştırmasında nitel veri analizi süreci için çoğunlukla insan uzmanlar tarafından gerçekleştirilen metin kodlama görevini de gerçekleştirme yeteneğine sahip olabilir. Nitel verileri kodlamak için uygulanan teknikler bulunmasına rağmen, şimdiye kadar hiçbir çalışma, özellikle pazarlama araştırması için kodlama görevlerinde GenAI araçlarını kullanmak için sistematik bir çerçeve sunmamıştır. Bu araştırma boşluğunu doldurmak için, mevcut çalışma 6 adımlı GenAI tabanlı bir içerik kodlama çerçevesi önermektedir. Çerçevede öncelikle marka mesajları toplanmakta ve sınıflandırma için ihtiyaç duyulan temalar tespit edilmektedir. Ardından marka mesajlarını arzu edilen temalara göre ayırmak için uygun Gen-AI modelleri seçilmekte, prompt hazırlanmakta, AI çıktıları olan sınıflar kodlanmakta ve son olarak kodlayıcılar arası uyuma bakılmaktadır. Bu doğrultuda, önerilen çerçeveyi test etmek için bir uygulama gerçekleştirilmiştir. Tüketici katılımı literatüründeki bilgilendirici, eğlendirici ve ödüllü içerik stratejileri tema olarak kullanılmış ve 3 Büyük Dil Modeli (LLM) ile insan uzmanlar arasındaki kodlama uyumları Kappa istatistikleri ile belirlenmiştir. Ulaşılan sonuçlara göre, İnsan ve ChatGPT arasındaki uyuşma düzeyi, bilgilendirici ve ödüllü içerikler için en iyi Kodlayıcılar Arası Uyum (IRR) vermiş olup Gemini, eğlenceli mesajlar için daha iyi performans göstermiştir. Varılan sonuçların önemli pratik katkıları arasında, pazarlama uygulayıcıları için yararlı ve daha hızlı kodlama süreçleri sunan mevcut çalışmanın önerilen çerçevesi öne çıkmaktadır.

Anahtar Kelimeler: *GenAI, büyük dil modeli, kodlayıcılar arası uyum, pazarlama, çerçeve.*

JEL Kodları: *M31, D83.*

"Bu çalışma Araştırma ve Yayın Etiğine uygun olarak hazırlanmıştır."

[2] Genişletilmiş Türkçe Özet, makalenin sonunda yer almaktadır.

1. INTRODUCTION

Since the mid-20th century, the need to process an increasing amount of data and obtain meaningful results has been among the important industrial problems. The concept of AI, which was put forward with the argument of "thinking machines" at the Dartmouth conference held in 1956, has been a revolutionary beginning in the autonomous processing, analysis and inference of data. In fact, data produced by consumers and brands in industrial markets is processed with various algorithms today. Since GenAI applications based on these algorithms offer significant cost advantages, marketing managers can use them strategically (Huang and Rust, 2021) for generating text, audio, image video etc. (Fui-Hoon Nah et al., 2023). Among these programs, GenAI based Large Language Models (LLMs) have been prominent in the last few years and can increase productivity in various areas, both for individual and business purposes (Hadi et al., 2024). The idea that LLMs can be useful for research and analysis in addition to content production has begun to be accepted in recent studies. These studies have mostly focused on the task of manual coding and classification of textual data, which is frequently used in quantitative methods (e.g., Demir, 2023; Theelen et al., 2024).

Although it is possible to extract meaning from detailed examination of qualitative text data, the results need to be quantified. In this respect, coding plays a crucial role in organizing and making numerical and unstructured data understandable. There are two types of coding: manual and electronic. Among these processes, choosing electronic methods such as package program or software can make the analysis deeper and easier (Basit, 2003). Marketing communication stands out among the fields where content analysis and coding techniques are frequently used. Especially in consumer engagement (CE) studies, manual coding techniques is frequently applied to identify social media message content strategies (Koçak et al., 2024). In the electronic coding method, pioneering software such as LIWC (Pennebaker et al., 2015) can be used. Thus, language styles such as adverbs, pronouns and prepositions in the text (e.g., Labrecque et al., 2020; Pezzuti et al., 2021) or formal elements such as the number of words that may affect CE (e.g., Koçak et al., 2024) can be determined. One of the problems that make the coding process difficult in CE studies is the large amount of data to be coded. So, objective coding of thousands of messages to determine which messages of brands receive the most interaction may be difficult and tiring. In this context, the researcher can seek help from independent coders (Ashley and Tuten, 2015). During the research, the selection of coders and classification of messages by these coders in accordance with the coding manual are quite difficult and time-consuming tasks. Therefore, there is a need for methods that help to code messages quickly, especially for CE studies in the literature. The aim of the current study is to fill this research gap by creating an AI-based content coding framework.

Content coding is carried out within the scope of qualitative data analysis. Thus, the hidden intention in the content of the message can be revealed (Prasad, 2008; Gupta

et al., 2017). At the end of the coding process, the judgments of two independent coders are compared numerically and the level of agreement between them is determined by Cohen's Kappa statistic (Landis and Koch, 1977). Although it is important to choose coders from experts in the literature, recently GenAI-based LLMs such as ChatGPT have also been preferred as coders (e.g., Demir, 2023; Theelen et al., 2024). Apart from ChatGPT, there are also popular AI models in the market such as Claude by Anthropic, Google Gemini and LLaMa by Meta (Buono et al., 2024). However, no study has been examined the level of agreement between these LLMs so far. To fill this gap, the current study uses various LLMs to determine the level of agreement between them with Kappa statistics. In this context, the research questions of the study can be developed as follows:

RQ1: What is the inter-rater reliability between AI and human coders?

RQ2: What is the inter-rater reliability between AI coders?

Considering these research questions, the findings of the current study will contribute to the understanding of the potential applications of LLMs in the literature that examines brand-generated content. Also, this study aims to fill the gap in the literature, which has very limited methodological approaches on electronic content coding and classification. Although content coding is frequently used in many areas, especially in communication studies, human-centered coding processes can be time-consuming and expensive (Lacy et al., 2015). The main novelty of this study is to provide an effective content coding framework that will facilitate the time-consuming, costly and difficult processes in the coding task and to make strategic decision-making processes more effective for industry professionals and scholars. Therefore, the rest of this study is organized as follows: Section 2 describes the proposed AI-based research framework. The third section performs an application to test the applicability of the proposed framework. Then, the findings are concluded, and the theoretical and practical implications of the study are given. Finally, suggestions for future studies are presented.

2. AI-BASED CONTENT CODING FRAMEWORK

With the digital transformation in the last half century, AI-based LLM technologies have been exhibited revolutionary potential to be used in industrial markets. Indeed, these tools have begun to offer the potential for use in almost every stage of marketing, such as CE, customer service, personalization, and communication (Paul et al., 2023). Thus, the desired content can be produced with these tools in accordance with marketing purposes. However, it is also necessary to determine the effectiveness of the content produced in marketing research. One of the processes used for this purpose is coding (e.g., Cvijikj and Michahelles, 2013; Menon et al., 2019). In this context, AI promises to be useful in tedious and time-consuming coding processes. In the literature, a limited number of studies have used LLM (ChatGPT) as a coder (Demir, 2023; Theelen et al., 2024), and to date, no studies have been conducted on an AI-based coding framework for consumer and market research. To fill this research gap,

the current study proposes a 6-step framework shown in Figure 1. The first step of the framework begins with deriving data from brands' online channels such as social media, web pages, and blogs using application programming interface (API)-like techniques. Second phase involves the identification of content themes. The third step explains how to decide which AI models will be used for classification. The fourth step guides how to design prompt strategies. In the fifth step, the responses produced by the AI considering the prompts are coded with the suitable variable types (e.g., dummy variable) that are related to content theme. Finally, the level of reliability between human and AI is checked with a suitable statistic.

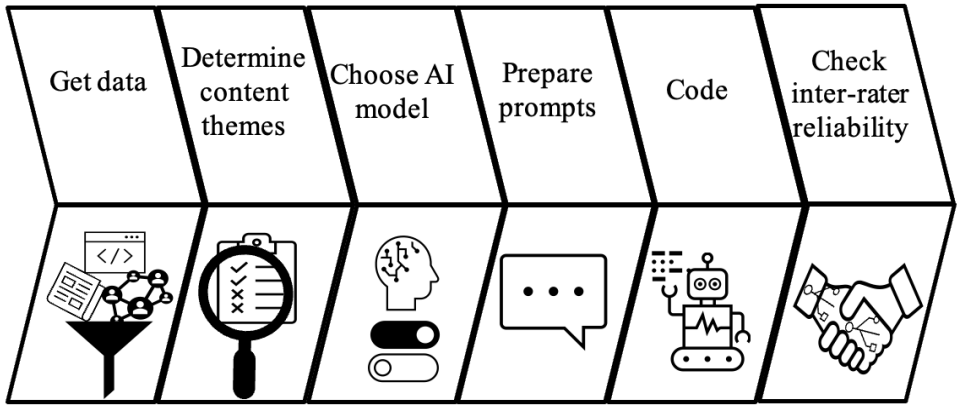

Figure 1. AI-based Content Coding Framework

1.1. Phase 1: Get data

Massive information in the online environment has the potential to affect not only content analysis processes but also researchers' approaches to data. In computer-aided solutions, large size of the data plays a very important role in the coding processes for content analysis (Lewis et al., 2013). Therefore, the first task of a researcher who will code with AI models is to access the data. Textual data created by brands or consumers in the online environment can be found on social media channels, websites, and news pages, and researchers who want to access this data can use API services, third-party applications, or scraping methods.

1.2. Phase 2: Determine content themes

At this stage of the framework, themes are identified through an in-depth literature review. Thus, a pre-arranged structure emerges for the classes to be coded. Content analysis is an applicable method for dividing content created by the brand or consumer into themes or classes. In this technique, themes need to be made understandable by the researcher to be classified reliably by different sources (Weber, 1990; Stemler, 2000).

1.3. Phase 3: Choose AI model

AI plays a vital role not only in task-based but also in daily search and use in society, and its performance in various applications makes it increasingly popular both academically and industrially. Among AI models, LLMs are undoubtedly remarkable and their use in society is increasing. Although ChatGPT stands out in these models, there are also LLMs that can show different capabilities in terms of performance for various tasks (Chang et al., 2024). Therefore, among the increasing number of LLMs in the current market for different or similar tasks, it is necessary to reach a conclusion about which one should be selected for the coding task. There are studies in the literature that technically evaluate LLMs in terms of their functional features (Chen et al., 2023). However, since the use of complex techniques can extend this stage of the framework, deciding which AI model to use by looking at the results of these studies can be a simple and effective solution. Moreover, indicators such as performance metrics, number of users, brand value or the position of the brand on the stock market can also be used at this stage.

1.4. Phase 4: Prepare prompts

To obtain the desired relevant and useful performance output from LLMs in the interaction processes, programing them with the right communication techniques is an advised task. Working with the principles of prompt engineering techniques which is a very necessary skill for information seekers or other users to get desirable answers from LLMs would also be useful for conversational AI researchers working in various industries such as education, health and security. Thus, the full potential of LLMs can be revealed (Marvin et al., 2023). This step of the study consists of two stages: Preparing coding manual and converting it into prompt. The coding manual is prepared in detail by considering the content themes identified in the second step of the proposed framework. Therefore, the coder easily distinguishes the meaning in the message content (Prasad, 2008; Gupta et al., 2017).

In the second step of the current phase, the prepared coding manual is converted into a prompt that AI can understand. Prompts are instructions given to an LLM to automate processes based on a set of rules and produce outputs of a certain quality. On the other hand, prompt engineering is defined as the skill set needed to communicate effectively with LLMs such as ChatGPT and Gemini (White et al., 2023). Recent studies have determined that there are many different prompt engineering techniques depending on the type and usage areas of LLM (Sahoo et al., 2024). Accordingly, the prompt engineering process basically consists of 5 steps. First, the purpose must be determined. Second, the capabilities of LLM including its capabilities and limitations must be understood. Third, the correct prompt format must be selected. Fourth, content related to the sought information must be produced, and finally, the received response must be tested and refined in line with the objectives (Marvin et al., 2023). In this step of the proposed framework, applying the correct

prompt technique is strongly required so that LLM can perform the desired coding. In this respect, using recent literature review studies (e.g., Sahoo et al., 2024) may be a useful approach.

1.5. Phase 5: Code

In this step of the framework, human and AI responses to the prepared coding manual and prompts are recorded. Content created by the brand, or the consumer is classified by coders according to their themes using a variable type such as dummy variable (1 vs 0).

1.6. Phase 6: Check inter-rater reliability

In the final step, the level of agreement between the human expert and the AI is calculated. In this stage, human acts as a controller during the coding process. In previous studies, some useful approaches such as Kappa (Cohen, 1960), IntraClass Correlation (ICC) coefficient (Gisev et al., 2013), Pearson's correlation coefficient, percentage of agreement and the Generalizability theory can be conducted for determining the Inter-Rater Reliability (IRR) between coders (Demir, 2023).

2. APPLICATION

The aim of the current study is to determine the usability of LLMs in CE studies. For this purpose, the coding made by humans and AI are compared with an empirical approach. Qualitative and quantitative methods are used together in this study. In this context, the study adopts descriptive approach (Siedlecki, 2020). At this stage of the study, Instagram messages published by the world's top 10 brands operating in different industries were selected using a simple random sampling method. The G*Power 3.1 program was used to calculate the sample size of the study (Faul et al., 2007). With an effect size of 0.5, a margin of error of 5%, and a power to represent the universe of 95%, the sample size was determined as 210. Therefore, a total of 300 messages from 10 brands were used for the proposed framework.

2.1. Data

The current study tests the applicability of the proposed framework using AI models as raters in the coding process. In this context, a total of 300 messages derived from the official Instagram pages of 10 brands (Amazon, Apple, AT&T, CocaCola, Disney, Google, Louis Vuitton, McDonald's, Toyota, and Visa) within the Kantar BrandZ 100 are analyzed. 30 messages from each brand were retrieved using Supermetrics[3] tool that is an extension on Google Sheets on August 11, 2024. During the compilation process, some Instagram metrics such as the creation time, post type (photo/video), content type, hashtags, number of messages, likes, comments and shares were obtained.

[3] https://supermetrics.com/ (Accessed: 11.08.2024)

2.2. Content themes

For the application of the current study, themes that may affect CE were identified with the guidance of past studies. In the CE literature, these themes are usually coded manually by the researcher. Brands may use informative, entertaining and remunerative contents on their social media pages. Previous studies have classified these messages using the manual coding method (e.g., Cvijikj and Michahelles, 2013; Menon et al., 2019). With this content coding method, researchers aim to analyze texts by revealing the hidden intentions behind the content created by brands (Prasad, 2008; Gupta et al., 2017). Therefore, in this study, coders were asked to code the message content types by considering whether they were informative, entertaining and/or remunerative following the literature (Ashley and Tuten, 2015).

2.3. AI model

A total of 3 AI-based LLM models were selected to encode the messages: ChatGPT (Brown et al., 2020), Gemini (Team et al., 2023) and Claude (Anthropic, 2024). Among the LLMs, trial versions of ChatGPT 4o and Gemini were used, and Claude purchased version.

2.4. Prompts

In this phase, prompt strategies based on the prepared coding manual were developed considering the themes identified in the second step. The coding manual prepared for the expert is given in Table 1. Informative, entertaining and remunerative content factors that may affect CE were arranged in this step, regarding previous studies (e.g., Cvijikj and Michahelles, 2013; Lee et al., 2018; Dolan et al., 2019).

Table 1. Coding Manual

Message content	Description	Tag
Informational	Message contains information about brands, companies, and products including news, prices, discounts, guarantees, release dates, locations, etc.	If the message has informative content, it will be coded as 1, otherwise as 0.
Entertaining	Message contains fun contents such as humor, emojis, sentences, slogans, wordplays, images (historic, scenic, occasion, celebrity, meme, animal etc.) and/or friendly communication.	If the message has entertaining content, it will be coded as 1, otherwise as 0.
Remunerative	Message contains remunerative contents such as rewards and benefits (e.g., economic incentives such as coupons and promotions)	If the message has remunerative content, it will be coded as 1, otherwise as 0.

2.4.1. Prompt manual for AI

In line with the purpose of the study, the coding manual was redesigned in a way that LLMs can understand. To obtain the desired outputs from LLMs, undeniable prompt engineering methods are essentially used. There are a variety of prompting techniques

to improve LLM performance, such as Zero-shot (Radford et al., 2019), few-shot (Brown et al., 2020), Chain-of-Thought (Wei et al., 2022) and Self-Consistency (Wang et al., 2022) (Sahoo et al., 2024). Moreover, various techniques can be applied separately for special tasks or academic purposes in the literature. Hybrid approaches can also be used in some cases (Giray, 2023). In the sample application of this study, standardized prompt strategies were developed for the classification of previously determined message content strategies by LLMs. Following the literature (Chen et al., 2023), the most suitable prompt technique was determined as Least-to-most prompting (Zhou et al., 2022) theoretically. The designed prompt strategies are tested on ChatGPT to get the desired outputs before the implementation, as seen in Figure 2.

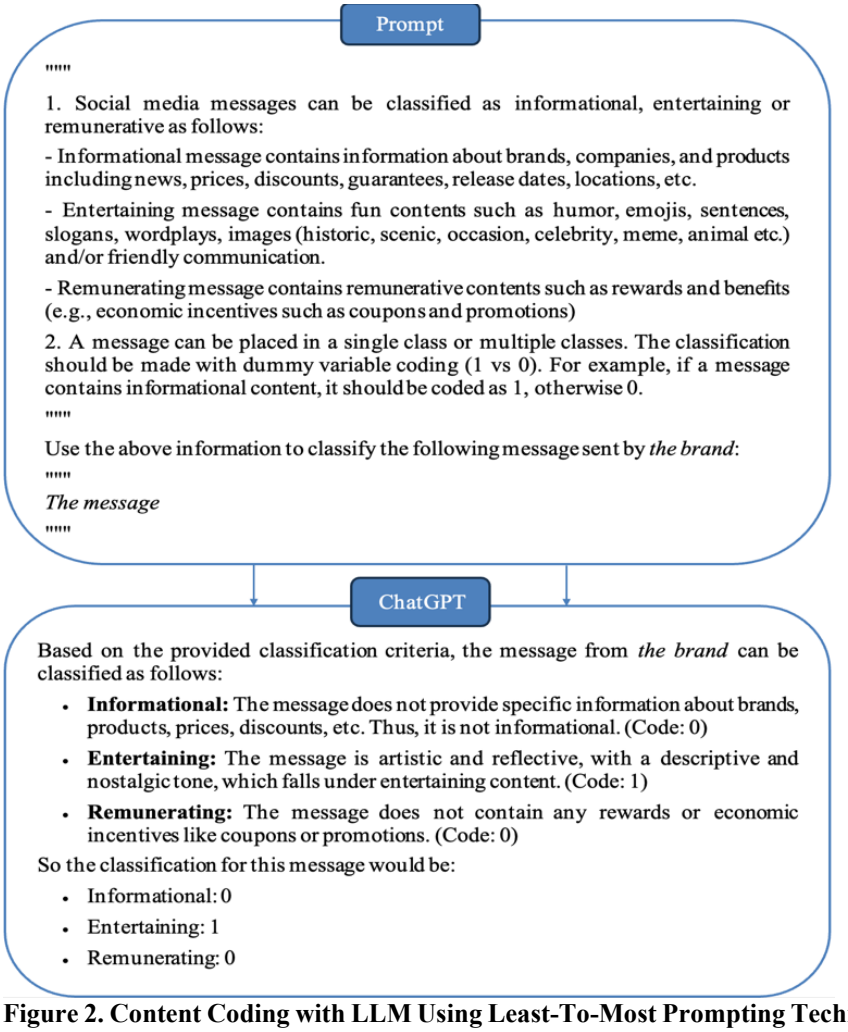

Figure 2. Content Coding with LLM Using Least-To-Most Prompting Technique

After checking the response that is obtained using Least-to-most prompting method, the LLM was asked to organize the response as a table. Thus, the output can be more visual, memorable and simple. This prompt was organized following the literature (Jin and Lu, 2023; Wang et al., 2024; Sahoo et al., 2024). The directives that were added to the prompt, shown in bold, and the outputs are shown in Figure 3.

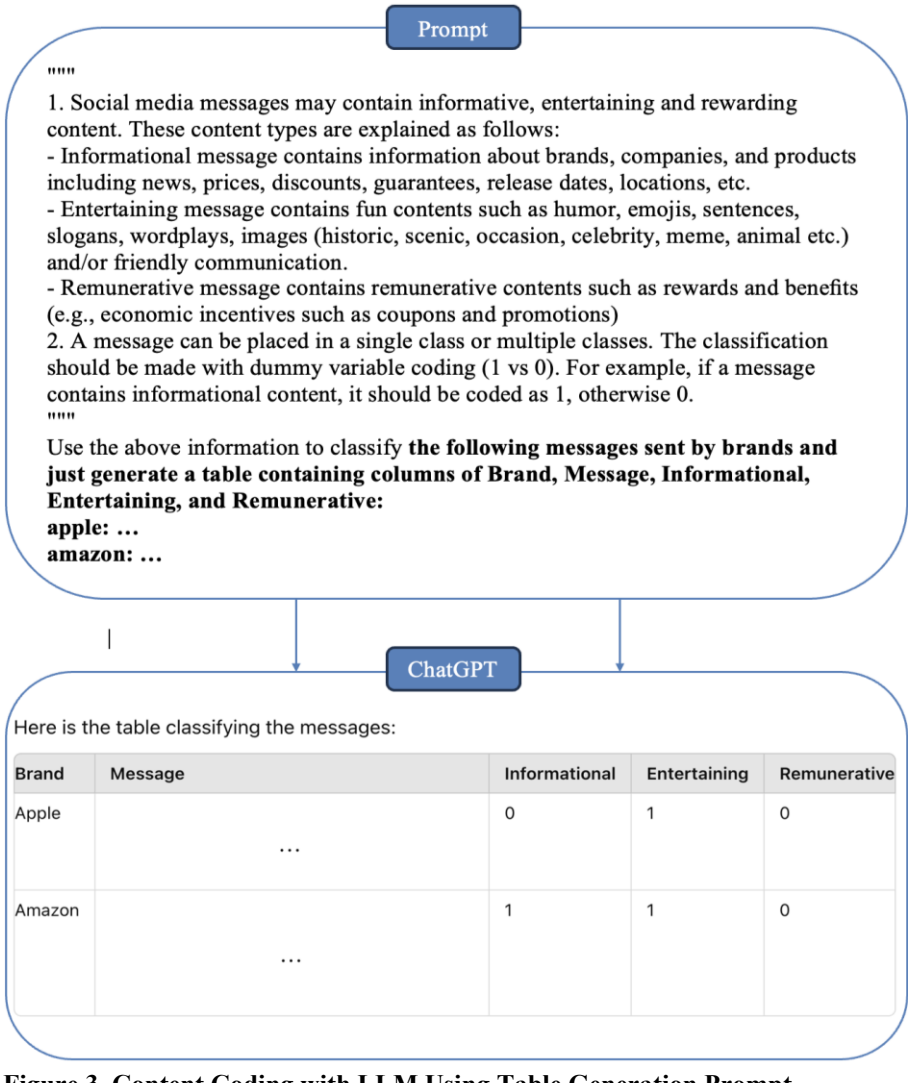

Figure 3. Content Coding with LLM Using Table Generation Prompt

A new instruction was added after the information in quotation marks for the LLM to create a table that shows brand message content strategies. After producing the table, the LLM provided detailed statistics for each brand. The instructions were rearranged

to prevent this situation, which increased the number of tokens. During the research, the coding times of both the expert and LLMs were compared, and it was observed that AI coding was much shorter. Another problem with LLMs is that none of them could code all messages. Therefore, the messages to be coded in the study were partly given to LLMs.

2.5. Coding

As a result of the applied coding and prompt methods, 300 Instagram messages belonging to brands were coded by experts and LLMs. Considering the relevant literature, a message may have more than one coding class (e.g., Ashley and Tuten, 2015; Koçak et al., 2024). Thus, no restriction has been made for LLMs and human raters in this regard (see Figure 3). According to the responses received, the messages were analyzed with the Excel program, respectively. Descriptive statistics including responses to brand messages are given in Table 2.

Table 2. Descriptives

Content strategy	Rater		ChatGPT-4o		Gemini		Claude 3.5 Sonnet		Mean	
	N	(%)	N	(%)	N	(%)	N	(%)		
Informational	182	60.6	152	50.6	179	59.6	154	51.3	167	55.53
Entertaining	264	88	210	70	275	91.6	266	88.6	254	84.55
Remunerative	16	5.3	13	4.3	51	17	32	10.6	28	9.3

When the content strategies on Instagram are examined, it is observed that brands mostly create entertaining messages (M=254, 84.55%). These messages are followed by informative (M=167, 55.53%) and remunerative (M=28, 9.3%) content, respectively.

2.6. Inter-rater reliability between human and AI

After coding content strategies, the level of agreement between experts and AI models was calculated using Kappa statistics. An effective way to check objective coding in research is to determine IRR. Kappa statistics measure the level of agreement for nominal variables and sets an IRR standard by correcting for agreement due to chance. In the method, the ratings of the two coders are cross tabulated for the observed agreement. Then, the agreement by chance is detected using the marginal frequencies of the responses and Kappa is calculated according to the following equation (Hallgren, 2012):

$$K = \frac{P_{(a)} - P_{(e)}}{1 - P_{(e)}}$$

where P_a represents the observed percentage of agreement, and P_e is the probability of expected agreement that is related to chance. Depending on the Kappa output values, agreement between coders is considered poor under 0, slight at 0.00-0.20, fair at 0.21-0.40, moderate at 0.41-0.60, substantial at 0.61-0.80, and perfect at 0.81-1.00. In this step of the sample application of the study, the AI and human coding were compared both within themselves and between each other using the Kappa statistic. The statistical findings of the comparisons are shown in Table 3. According to the findings obtained from the Human and AI comparison, when informational messages are considered, the Kappa statistics show that the best level of agreement is between the human expert and the ChatGPT-4o at the moderate level (K = 0.41). This score is followed by Gemini (K = 0.37) and Claude 3.5 Sonnet (K = 0.36) versions, respectively. Considering entertaining messages, Gemini gives the best score with the human coding at the fair level (K = 0.36). Human-ChatGPT and Human-Claude get equal scores (K = 0.16). For remunerative contents, Human-ChatGPT comparison received the best score with the substantial level (K = 0.75). This score is followed by Human-Claude (K=0.33) and Human-Gemini (K=0.27) at the fair level.

Table 3. IRR Reliability

Rater	Comparison	Kappa		
		Informational	Entertaining	Remunerative
AI - AI	ChatGPT-4o vs Gemini	0.50	0.23	0.33
	ChatGPT-4o vs Claude3.5	0.61	0.42	0.36
	Gemini vs Claude3.5	0.74	0.27	0.65
Human - AI	Human vs ChatGPT-4o	0.41	0.16	0.75
	Huma vs Gemini	0.37	0.36	0.27
	Human vs Claude3.5	0.36	0.16	0.33

When considering informational content coding, Gemini vs Claude (K=0.74) took the best score among AI models at the substantial level. This was followed by ChatGPT vs Claude (K=0.61) at the substantial level and ChatGPT vs Gemini (K=0.50) at the moderate level, respectively. Among the entertaining content, ChatGPT vs Claude (K=0.42) received the best Kappa score at the moderate level, followed by Gemini vs Claude (K=0.27) and ChatGPT vs Gemini (K=0.23) scores at the fair level. Finally, looking at the remuneration content coding among AI models, Gemini vs Claude (K = 0.65) got the best score at the substantial level, followed by ChatGPT vs Claude (K = 0.36) and ChatGPT vs Gemini (K = 0.33) at the fair level, respectively.

3. CONCLUSION AND DISCUSSION

As a results of recent developments in AI modeling, LLMs have led significant potential not only for academic research but also for daily tasks such as content generation. To produce valuable insights into human behavior, especially in social and cultural terms, these tools are used in various disciplines such as health and education (Demir, 2023). Moreover, they provide new avenues for utilizing qualitative research methods commonly applied in the field (Theelen et al., 2024).

However, within the scope of detailed literature review, no study has visually presented an AI-based coding framework for marketing research to date. In this direction, the current study has proposed a 6-step AI-based framework that can be used for coding needs in the field. A total of three important content strategy (informative, entertaining and remunerating) in CE literature were selected as a theme in the framework. Then, they were coded by human and AI, respectively and the level of agreement between coders were compared using Kappa statistics. According to the results of application, ChatGPT gave the best results for informative and remunerative messages in the Human-AI comparison. Gemini showed better performance for entertaining messages. The level of agreement between AI and AI had better Kappa score than human and AI. A possible explanation for this situation is that both LLMs may use the same training data. Accordingly, Gemini and Claude 3.5 Sonnet obtained the best score for informational and remunerative message coding. ChatGPT-4o and Claude 3.5 Sonnet got better Kappa statistic for entertaining messages. Although the level of agreement between human and ChatGPT was good in remunerative content, it is not at the desired levels for other themes. However, the findings are promising for future research.

4. THEORETICAL AND PRACTICAL IMPLICATIONS

In terms of the proposed framework and the results of application, the current study also offers some theoretical and practical contributions to the literature and industry. The first and major theoretical contribution is that the current study addresses how to classify contents in marketing area. Although there are some programs to classify content automatically according to the structures of the messages (Pennebaker et al., 2015), there is a need for a systematic framework to be used in classifying marketing content according to their subjects. Therefore, this study proposes an AI-based coding framework for the classification task. Second, this study compared the judgements of human expert and three popular LLMs. Looking at the literature, few studies used ChatGPT as a coder (Demir, 2023; Theelen et al., 2024), but no study has made a comparison between human and AI for the classification of marketing content. In this respect, this study uses Kappa statistics to test the applicability of the proposed framework.

The current study also makes valuable practical implications for brands, marketing managers and academic researchers. First, most of studies in the LLMs literature proposes and uses suitable prompting strategies to uncover the real potential of AI tools (e.g., Giray, 2023; Chen et al., 2023). To boost the applicability of the proposed framework, this study applied some useful prompting techniques for the coding manual to get relevant answers from LLMs. Thus, coding processes can be carried out faster. Second, AI tools can facilitate marketing processes with the social interaction, communication practices and cost-effective structures they offer (Paul et al., 2023). In this context, GenAI applications can be used especially in the classification of large volumes of textual data (Ollion et al., 2023). Therefore, the proposed framework of the current study also eliminates the costs arising from the need for human experts. Third, to measure whether the text is classified objectively or not during the coding

process, checking the agreement between coders quantitatively is an important task (Demir, 2023). In this regard, Kappa is a useful statistic to measure the level of agreement (Landis and Koch, 1977). The findings of this study suggest that GenAI tools can classify only informational and remunerative contents with high level of agreement. Therefore, it would be better for GenAI developers to train GenAI tools to understand metaphorical and/or entertaining content.

5. LIMITATIONS AND FUTURE RESEARCH

Despite all these theoretical and practical implications, this study has some limitations that will give research directions for future studies. The first limitation of the study is that a total of 3 AI-based LLMs, ChatGPT, Gemini and Claude, were used in the study because they were more easily accessible. According to recent studies, there are many types of LLMs (Minae et al., 2024). Therefore, future studies should investigate the level of agreement between different LLMs. Second, three basic factors affecting CE were considered for the coding process in the application. However, previous studies determined that figurative elements (Koçak and Atalık, 2024), grammar types and language structures (Koçak, 2021; Labrecque et al., 2020; Koçak, 2023) could affect CE. Thus, future studies should examine different themes in coding process. Finally, CE literature in marketing was applied to test the proposed framework. LLMs are used not only in marketing but also in different disciplines such as health and education (Demir, 2023; Theelen et al., 2024). Thus, testing the proposed framework for different disciplines in future research will provide both theoretical contributions to the literature and practical contributions to the industry.

PAZARLAMADA YAPAY ZEKAYI DEĞERLENDİRİCİ OLARAK NASIL KULLANIRIM? ÜRETKEN AI İLE KAPSAMLI BİR İÇERİK KODLAMA ÇERÇEVESİ

1. GİRİŞ

Markalar tarafından üretilen veriler günümüzde çeşitli algoritmalarla işlenmekte ve bu algoritmalara dayalı GenAI uygulamaları önemli maliyet avantajları sağladığından pazarlama yöneticileri bunları stratejik olarak (Huang ve Rust, 2021) metin, ses, görüntü video vb. üretmek için kullanabilmektedirler (Fui-Hoon Nah vd., 2023). Son yıllarda öne çıkan GenAI tabanlı Büyük Dil Modelleri (BDM) ise hem bireysel hem de ticari amaçlar için çeşitli alanlarda üretkenliği artırabilmektedir (Hadi vd., 2023). BDM'lerin içerik üretiminin yanı sıra araştırma ve analiz için de yararlı olabileceği fikri son çalışmalarda kabul görmeye başlamıştır. Bu çalışmalar çoğunlukla nicel yöntemlerde sıklıkla kullanılan metinsel verilerin manuel kodlanması ve sınıflandırılması görevine odaklanmıştır (Demir, 2023; Theelen vd., 2024).
Alanyazında manuel ve elektronik olmak üzere iki tür kodlama biçimi bulunmaktadır. Özellikle tüketici katılımı (TK) çalışmalarında, sosyal medya mesaj içerik stratejilerini belirlemek için manuel kodlama teknikleri sıklıkla uygulanmaktadır (Koçak vd., 2023). Elektronik kodlama yönteminde ise LIWC (Pennebaker vd., 2015) gibi öncü yazılımlar kullanılmaktadır. İçerik kodlaması nitel veri analizinin önemli bir parçasıdır. Bu sayede mesaj içeriğindeki gizli niyet ortaya çıkarılabilir (Prasad, 2008; Gupta vd., 2017). Kodlama süreci sonunda iki bağımsız kodlayıcının yargıları sayısal olarak karşılaştırılır ve aralarındaki uyuşma düzeyi Cohen'in Kappa istatistiği ile belirlenir (Landis ve Koch, 1977). Kodlayıcıların literatürde uzman kişilerden seçilmesi önemli olmakla birlikte son dönemde ChatGPT de kodlayıcı olarak tercih edilmektedir (örn. Demir, 2023; Theelen vd., 2024). ChatGPT dışında piyasada Anthropic tarafından Claude, Google Gemini ve Meta tarafından LLaMa gibi popüler AI modelleri de bulunmaktadır (Buono vd., 2024). Ancak, literatürde şimdiye dek üretken-AI araçları arasındaki uyum düzeyini inceleyen hiçbir çalışma yapılmamıştır. Mevcut çalışma önerdiği çerçeve ile bahsi geçen boşluğu doldurmayı amaçlamaktadır.

2. ÖNERİLEN ÇERÇEVE

Mevcut çalışmanın önermiş olduğu 6 adımlı çerçevenin ilk adımında, uygulama programlama arayüzü (API) benzeri teknikler kullanılarak markaların sosyal medya, web sayfaları ve bloglar gibi çevrimiçi kanallarından veri türetilmektedir. İkinci aşamada, içerik temaları tanımlanmakta; üçüncü adım ise sınıflandırma için hangi AI modellerinin kullanılacağı kararını içermektedir. Dördüncü adım, istem (prompt) stratejilerinin nasıl tasarlanacağına rehberlik ederken beşinci adımda, AI tarafından üretilen yanıtlar, içerik temasıyla ilişkili değişken türleriyle (örneğin, kukla değişken) kodlanmaktadır. Son olarak, insan ve AI arasındaki güvenilirlik düzeyi Kappa gibi uygun bir istatistikle kontrol edilmektedir.

3. UYGULAMA

Mevcut çalışmanın uygulamasında insan ve yapay zekâ tarafından yapılan kodlamalar ampirik bir yaklaşımla karşılaştırılmıştır. Önerilen çerçevenin ilk adımında Kantar BrandZ 100 içindeki 10 markanın (Amazon, Apple, AT&T, CocaCola, Disney, Google, Louis Vuitton, McDonald's, Toyota ve Visa) resmi Instagram sayfalarından toplam 300 mesaj analiz edilmiştir. Çerçevenin ikinci adımında ilgili literatürden yola çıkılarak mesajlar bilgilendirici, eğlendirici ve ödüllü içerikler olarak temalandırılmıştır. Çerçevenin sonraki adımında mesajların ilgili temalar dikkate alınarak kodlanması için günümüzün trend AI tabanlı BDM modellerinden olan ChatGPT (Brown vd., 2020), Gemini (Team vd., 2023) ve Claude (Anthropic, 2024) kullanılmıştır. Çerçevenin dördüncü adımında BDM'lerden istenen çıktıları elde etmek için gerekli prompt mühendisliği yöntemleri kullanılmıştır. En azdan en çok istem yöntemi dikkate alınarak elde edilen AI yanıtlarının literatüre (Jin ve Lu, 2023; Wang ve ark., 2024; Sahoo ve ark., 2024) uygun olarak tablo halinde düzenlenmesi istenmiştir. Çerçevenin beşinci adımında markalara ait 300 Instagram mesajı uzman ve BDM'ler tarafından kodlanmış olup Instagram'daki içerik stratejileri incelendiğinde markaların çoğunlukla eğlenceli mesajlar paylaştığı tespit edilmiştir. Çerçevenin son adımında ise AI modelleri arasındaki uyuşma düzeyi Kappa istatistiği kullanılarak hesaplanmıştır. İnsan ve AI karşılaştırmasından elde edilen bulgulara göre bilgilendirici mesajlar ele alındığında Kappa istatistiği en iyi uyuşma düzeyinin insan uzman ve ChatGPT-4o arasında olduğu gözlenmiştir. Eğlenceli mesajlar ele alındığında Gemini en iyi puanı almıştır. Ödüllü içerikler için ChatGPT daha iyi sonuç üretmiştir.

4. SONUÇ

İnsan davranışına, özellikle sosyal ve kültürel açıdan, değerli içgörüler üretmek için AI araçları, sağlık ve eğitim gibi çeşitli disiplinlerde kullanılmaktadır (Demir, 2023). Literatür incelendiğinde son derece kısıtlı sayıda çalışma ChatGPT'yi kodlayıcı olarak kullanmış olup (Demir, 2023; Theelen ve diğerleri, 2024) mevcut çalışmada önerilen çerçevenin uygulanabilirliğini test etmek için insan uzman ve üç üretken-AI kodlayıcı olarak kullanılmıştır. TK literatüründeki üç önemli içerik stratejisi (bilgilendirici, eğlenceli ve ödüllendirici) çerçevede bir tema olarak seçilmiş ve bunlar sırasıyla insan ve AI tarafından kodlanarak kodlayıcılar arasındaki uyum düzeyi, Kappa istatistikleri dikkate alınarak karşılaştırılmıştır. İnsan ve ChatGPT arasındaki uyum düzeyi, ödüllendirici içerikte iyi olsa da diğer temalar için istenen düzeyde değildir. Ancak bulgular gelecekteki araştırmalar için ümit vericidir.

Uygulama sonuçları açısından, önerilen çerçeve endüstri, pazarlama yöneticileri ve akademik araştırmacılar için uygulanabilirdir. Çalışmanın diğer bir pratik katkısı, kodlama kılavuzu için hazırlanan prompt stratejisidir. Böylece kodlama süreçleri daha hızlı ve daha az maliyetle gerçekleştirilebilmektedir.

REFERENCES

Anthropic. (2024). Claude 3.5 Sonnet. Accessed: 7.30.2024. https://www.anthropic.com/news/claude-3-5-sonnet

Ashley, C., and Tuten, T. (2015). Creative strategies in social media marketing: An exploratory study of branded social content and consumer engagement. *Psychology & Marketing*, 32(1), 15-27.

Basit, T. (2003). Manual or electronic? The role of coding in qualitative data analysis. *Educational research*, 45(2), 143-154.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

Buono, D., Felecan, M., and Tessitore, C. (2024). An introduction to Large Language Models and their relevance for statistical offices. Eurostat: Luxembourg. Accessed: 27.08.2024 https://ec.europa.eu/eurostat/documents/3888793/18771440/KS-TC-24-001-EN-N.pdf/fbdbcc5b-7b93-39af-5980-944112feaff6?version=1.0&t=1711031922718

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1-45.

Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2023). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

Cvijikj, I. P., and Michahelles, F. (2013). Online engagement factors on Facebook brand pages. *Social Network Analysis and Mining*, 3(4), 843–861. https://doi.org/10.1007/s13278-013-0098-8

Demir, S. (2023). Investigation of ChatGPT and Real Raters in Scoring Open-Ended Items in Terms of Inter-Rater Reliability. *Uluslararası Türk Eğitim Bilimleri Dergisi*, *11*(21), 1072-1099. https://doi.org/10.46778/goputeb.1345752

Dolan, R., Conduit, J., Frethey-Bentham, C., Fahy, J., and Goodman, S. (2019). Social media engagement behavior: A framework for engaging customers through social media content. *European journal of marketing*, 53(10), 2213-2243.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., and Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3), 277-304.

Giray, L. (2023). Prompt engineering with ChatGPT: a guide for academic writers. *Annals of biomedical engineering*, *51*(12), 2629-2633.

Gisev, N., Bell, J. S., and Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3), 330-338.

Gupta, H., Singh, S., and Sinha, P. (2017). Multimedia tool as a predictor for social media advertising-a YouTube way. *Multimedia tools and applications*, 76(18), 18557-18568.

Hadi, M. U., Al Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M., ... & Shah, M. (2024). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*. https://doi.org/10.36227/techrxiv.23589741.v4.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1), 23-34. doi: 10.20982/tqmp.08.1.p023

Huang, M. H., and Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, 49, 30-50.

Jin, Z., and Lu, W. (2023). Tab-cot: Zero-shot tabular chain of thought. *arXiv preprint arXiv:2305.17812*.

Koçak, B. B. (2021). Fly "With us"! Impact of consumer-brand relationship on consumer engagement: An empirical investigation on Turkish airline instagram pages. *Tüketici ve Tüketim Araştırmaları Dergisi*, 13(2), 253–282. https://doi.org/10.15659/ttad.13.2.139

Koçak, B. B. (2023). Impact of Brand Linguistic Characteristics on Consumer Engagement: A Psycholinguistics Approach for Airline Facebook Pages. M. Dalkılıç (Ed.), *INSAC 2023 New Trends in Social and Education Sciences* içinde, (13-30). Ankara: Duvar.

Koçak, C. B., and Atalık, Ö. (2024). Figurative language effect on consumer engagement: an empirical investigation for Turkish airline industry. *Aviation*, 28(2), 128-140.

Koçak, C. B., Atalık, Ö., and Koçak, B. B. (2024). Mecazi dil unsurlarının tüketici katılımı üzerindeki etkisi: Türk havayolu Instagram sayfaları örneği. *Pazarlama ve Pazarlama Araştırmaları Dergisi*, 17(1), 1-38.

Labrecque, L. I., Swani, K., and Stephen, A. T. (2020). The impact of pronoun choices on consumer engagement actions: Exploring top global brands' social media communications. *Psychology & Marketing*, 37(6), 796-814.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & mass communication quarterly*, *92*(4), 791-811.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Lee, Dokyun, Kartik Hosanagar, and Harikesh Nair (2018), "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook," *Management Science*, 64 (11), 5105–31.

Lewis, S. C., Zamith, R., and Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of broadcasting & electronic media*, 57(1), 34-52.

Marvin, G., Hellen, N., Jjingo, D., and Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics* (pp. 387-402). Singapore: Springer Nature Singapore.

Menon, R. V., Sigurdsson, V., Larsen, N. M., Fagerstrøm, A., Sørensen, H., Marteinsdottir, H. G., and Foxall, G. R. (2019). How to grow brand post engagement on Facebook and Twitter for airlines? An empirical investigation of design and content factors. *Journal of Air Transport Management*, 79, Article 101678. https://doi.org/10.1016/j.jairtraman.2019.05.002

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Ollion, E., Shen, R., Macanovic, A., & Chatelain, A. (2023). ChatGPT for Text Annotation? Mind the Hype!. https://fles.osf.io/v1/resources/x58kn/providers/osfstorage/651d60731bc8650a79f376cf?direct=&mode=render.

Paul, J., Ueno, A., and Dennis, C. (2023). ChatGPT and consumers: Benefits, pitfalls and future research agenda. *International Journal of Consumer Studies*, 47(4), 1213-1225.

Prasad, B. D. (2008). Content analysis. *Research methods for social work*, *5*(le20).

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. Austin, TX: University of Texas at Austin.

Pezzuti, T., Leonhardt, J. M., and Warren, C. (2021). Certainty in language increases consumer engagement on social media. *Journal of Interactive Marketing*, 53(1), 32-46.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Siedlecki, S. L. (2020). Understanding descriptive research designs and methods. *Clinical Nurse Specialist*, 34(1), 8-12.

Stemler, S., (2000) "An overview of content analysis", *Practical Assessment, Research, and Evaluation* 7(1): 17. doi: https://doi.org/10.7275/z6fm-2e34

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Theelen, H., Vreuls, J., and Rutten, J. (2024). Doing Research with Help from ChatGPT: Promising Examples for Coding and Inter-Rater Reliability. *International Journal of Technology in Education*, *7*(1), 1-18.

Wang, Z., Zhang, H., Li, C. L., Eisenschlos, J. M., Perot, V., Wang, Z., ... & Pfister, T. (2024). Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Weber, R. P. (1990). Basic Content Analysis, 2nd ed. Newbury Park, CA.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., ... & Chi, E. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

KATKI ORANI / CONTRIBUTION RATE	**AÇIKLAMA / EXPLANATION**	**KATKIDA BULUNANLAR / CONTRIBUTORS**
Fikir veya Kavram / Idea or Notion	Araştırma hipotezini veya fikrini oluşturmak / Form the research hypothesis or idea	Bahri Baran KOÇAK
Tasarım / Design	Yöntemi, ölçeği ve deseni tasarlamak / Designing method, scale and pattern	Bahri Baran KOÇAK
Veri Toplama ve İşleme / Data Collecting and Processing	Verileri toplamak, düzenlenmek ve raporlamak / Collecting, organizing and reporting data	Bahri Baran KOÇAK
Tartışma ve Yorum / Discussion and Interpretation	Bulguların değerlendirilmesinde ve sonuçlandırılmasında sorumluluk almak / Taking responsibility in evaluating and finalizing the findings	Bahri Baran KOÇAK
Literatür Taraması / Literature Review	Çalışma için gerekli literatürü taramak / Review the literature required for the study	Bahri Baran KOÇAK