Hochschule Neu-Ulm
University of Applied Sciences

Master Thesis
in the master program
Artificial Intelligence and Data Analytics
at University of Applied Sciences Neu-Ulm

Emotional Impact of AI-Generated Content: Exploring Emotion and Engagement on Social Media

1st examiner Prof. Dr. Andy Weeger

2nd examiner: Prof. Dr. Marten Risius

Author: Cedric Bretzinger (Enrolment number: 277201)

Topic received: 01.01.2025

Date of submission: 18.06.2025

Abstract

Generative-AI imagery is flooding social media timelines, hence platforms and regulators now urge creators to flag such work as AI-generated. Whether that transparency clarifies or distorts user experience remains contested: lab studies point to suspicion and diminished appeal, whereas field anecdotes highlight excitement and virality. We add to that debate with a large-scale naturalistic study of Instagram data. A computational content analysis is conducted on 64 806 comments drawn from top-ranked posts in the #aiart and #traditionalart hashtags. A dual-stage NLP pipeline, zero-shot GPT labelling for 62 languages and a GoEmotions-fine-tuned BERT model for English extracts discrete emotions, sentiment, and a four-level measure of engagement depth. Findings reveal a consistent trade-off: When AI authorship is disclosed admiration and other high-esteem emotions recede while disapproval and mild amusement become more common. Yet the same posts draw noticeably more likes, comments, and reshares while the comments themselves tend to be brief and unreflective. In English-language the cooling of positive affect and surge of disapproval are most pronounced.

Taken together, the results suggest that provenance tags work less as neutral information and more as affective signals: they dampen perceptions of authenticity despite the novelty of algorithmic art boosting superficial engagement. The study extends transparency and algorithm-aversion research from controlled settings into a real social-media environment and offers practical guidance for platform design and policy.

Keywords: generative AI, transparency, Instagram, emotion analysis, user engagement

Table of Contents

List of Abbreviations ..... **VI**

List of Figures ..... **VII**

List of Tables ..... **VIII**

1 Introduction **1**

2 Literature Review **3**

2.1 Foundations of Artificial Intelligence in Creative Contexts 3

2.2 Transparency and Content Origin on Social Media 5

2.3 Theoretical Foundations of Emotion 7

2.4 Emotion in Response to AI and Digital Artifacts 9

2.5 Algorithm Aversion and Trust in AI Systems 10

2.6 User Engagement on Social Media Platforms 12

2.7 Emotion Detection in Text: Capabilities and Methods 13

2.8 Research Gap and Thesis Positioning 15

3 Research Model **16**

3.1 Content Origin and Affective Response (H1, H2) 16

3.2 Content Origin and Engagement Behavior (H3, H4) 16

3.3 Moderating Role of User Exposure Pattern (H5) 17

3.4 Transparency as a Control Factor 17

3.5 Conceptual Research Model 18

4 Research Methodology **19**

4.1 Research Design 19

4.2 Philosophical Positioning 20

4.3 Methodological Justification in Information Systems Research 20

4.4 Data Source and Collection Procedure 21

4.5 Emotion Classification Pipeline 23

4.5.1 GoEmotions Taxonomy 24

4.5.2 ChatGPT for Weak Labeling 26

4.5.3 BERT-Based Emotion Detection 27

4.5.4 Label Harmonization and Source Attribution 28

4.6 Sentiment Classification Approach 29

4.7 Engagement Metrics 29

4.8 Engagement Depth Classification 30

4.9 User Segmentation by Exposure Pattern 32

4.10 Variable Operationalization 32

4.11 Statistical Analysis 34

4.11.1 Emotion Distribution (H1) 34

4.11.2 Sentiment Polarity (H2) 35

4.11.3 Engagement Volume (H3) 36

4.11.4 Engagement Depth (H4) 36

4.11.5 User Exposure Pattern (H5) 37
4.11.6 Summary of Statistical Procedures 38
4.11.7 Overview of Analytical Environment 38
4.12 Methodological Limitations 39
4.12.1 Data Collection and Sampling 39
4.12.2 Measurement and Annotation 39
4.12.3 Model Assumptions and Statistical Inference 40
4.12.4 External Validity and Interpretation 40
4.12.5 Technical Constraints 40
4.13 Ethical Considerations 41
4.14 Methodological Reflections and Trade-offs 41
5 Results 43
5.1 Descriptive Summary of the Dataset 43
5.2 Classifier Concordance and Validation 44
5.2.1 Plausibility Audit of Classifier Outputs 44
5.2.2 Agreement Between GPT and BERT Classifiers 45
5.2.3 Consistency of Sentiment Polarity Derived from Emotion Labels 46
5.3 Emotional Response (H1) 46
5.3.1 Model Summary and Screening Procedure 46
5.3.2 Emotion-Specific Effects: Odds Ratios and Predicted Probabilities 47
5.3.3 Effects in the English Subset 47
5.3.4 Visual Comparison: Full Sample vs. English Subset 48
5.4 Sentiment Polarity (H2) 49
5.4.1 Distribution of Positive, Neutral, and Negative Labels 49
5.4.2 Chi-Square Test and Effect Size 50
5.4.3 Standardized Residuals and Pairwise Differences 50
5.5 Behavioral Engagement (H3) 51
5.5.1 Like Count Analysis 52
5.5.2 Comment Count Analysis 52
5.5.3 Reshare Count Analysis 52
5.5.4 Engagement Summary 53
5.6 Depth of Engagement (H4) 53
5.6.1 Descriptive Distribution and Contingency Analysis 54
5.6.2 Standardized Residuals and Pairwise Comparison 54
5.6.3 Multinomial Logit Model 55
5.7 Moderating Role of Exposure Pattern (H5) 55
5.7.1 Interaction Effects by Exposure Group 55
5.7.2 Emotional Probabilities Across Exposure Groups 56
5.8 Summary of Findings 56
6 Discussion 57
6.1 Implications for Theory 57
6.1.1 Transparency Cues as Signals and Affective Triggers 57
6.1.2 Volume Without Depth: An Engagement Paradox 58
6.1.3 Cultural Appraisal and Emotional Amplification 58
6.1.4 Familiarity Without Reappraisal 59
6.1.5 Summary: Affective Structures and Methodological Insights 59
6.2 Implications for Practice and Policy 59
6.3 Study Limitations and Boundary Conditions 60
6.3.1 Sampling Frame and Platform Ecology 61
6.3.2 Cross-Sectional Snapshot 61
6.3.3 Hashtag-Based Origin Coding and Potential Misclassification 61
6.3.4 Emotion Labeling Accuracy and Taxonomic Limitations 61

6.3.5 Nested Data and Unmodelled Dependence 61
6.3.6 Cultural and Linguistic Generalizability 61
6.3.7 Ethical Constraints and Technical Risks 61
6.3.8 Statistical Modeling Constraints 62
6.3.9 External Validity and Platform-Specific Interpretive Scope 62
6.3.10 Subjectivity in Human Audit 62
6.3.11 Authenticity of User Accounts 62
6.4 Future Research Directions 62

7 Conclusion 65

References 67

Appendix A: Declaration on the use of GenAI tools 1

Appendix B: Final LLM instruction prompt 2

Appendix C: Requirements 4

Appendix D: Summarized distributions of classifier confidence 5

Appendix E: Full Emotion Effects 7

Appendix F: English-Subset Emotion Effects 8

Appendix G: Pairwise chi-square tests between engagement depth levels 9

List of Abbreviations

APPs ..... *average predicted probabilities*
BERT ..... *Bidirectional Encoder Representations from Transformers*
C2PA ..... *Coalition for Content Provenance and Authenticity*
CI ..... *confidence interval*
CSS ..... *computational social science*
DEQ ..... *Discrete Emotions Questionnaire*
GDPR ..... *General Data Protection Regulation*
GenAI ..... *Generative Artificial Intelligence*
GPT ..... *Generative Pre-trained Transformer*
IRR ..... *incidence-rate ratio*
IS ..... *information systems*
ISR ..... *Information Systems Research*
LLMs ..... *Large Language Models*
MNL ..... *multinomial logit*
NBR ..... *Negative Binomial Regression*
NLP ..... *natural language processing*
OR ..... *odds ratio*
SOTA ..... *state-of-the-art*
UGT ..... *Uses and Gratifications Theory*

List of Figures

Figure 1: Generative AI and other AI concepts, adapted from Banh and Strobel (2023, p. 2), inspired by Goodfellow et al. (2016, p. 9) and Janiesch et al. (2021, p. 687) 3

Figure 2: Risk-based pyramid from the EU AI Act, adopted from European Parliament and Madiega (2024, p. 8) 5

Figure 3: Research Model. *Content Origin is coded only for posts whose artificial origin is transparently disclosed. 18

Figure 4: Data Collection and Storage Pipeline, *including Anonymization of UserID 21

Figure 5: Text Classification Pipeline 23

Figure 6: Hierarchical clustering of GoEmotions categories by sentiment orientation and semantic proximity, adapted from Demszky et al. (2020, p. 5) 25

Figure 7: Distribution of annotated emotion categories in the GoEmotions Reddit dataset, adapted from Google Research (2021) 26

Figure 9: Snippet of Emotion Classification Prompt 27

Figure 10: Snippet of Engagement Depth Classification Prompt 31

Figure 10: Language distribution by content origin 44

Figure 11: Confusion Matrix (GPT rows × BERT columns) 45

Figure 12: Confusion Matrix for Sentiment Polarity Agreement 46

Figure 13: Predicted Probability Differences (Δ(APP)) by Emotion and Dataset 49

Figure 14: Sentiment distribution by content origin 50

Figure 15: Standardized residuals for Origin × Sentiment 51

Figure 16: Incidence-rate ratios (IRR) for likes, comments, and reshares 53

Figure 17: Model-based predicted engagement volumes by content origin 53

Figure 18: Engagement depth distribution by content origin 54

List of Tables

Table 1: Hypotheses Overview 18
Table 2: Methodology trends over the years 2013-2018, adapted from Mazaheri et al. (2020, p. 12) 21
Table 3: Raw Data Fields Structure 22
Table 4: Few-Shot Comment per Engagement Depth Category 31
Table 5: Operationalization of Variables 34
Table 6: Statistical Procedures 38
Table 7: Core Libraries and Packages 38
Table 8: Dataset Overview 43
Table 9: Content Origin Distribution 43
Table 10: Human Audit: Agreement Metrics for GPT and BERT 44
Table 11: Model Summary Statistics for Multinomial Logit Regression on Emotion Labels 47
Table 12: Emotion-specific effects of AI origin relative to human content 47
Table 13: Emotion-Specific Effects of AI Origin in the English Subset 48
Table 14: Frequency of sentiment categories by content origin (GPT-classified) 50
Table 15: Standardized residuals from the contingency table 50
Table 16: Pairwise chi-square contrasts between sentiment categories 51
Table 17: Predicted Mean Likes by Content Origin 52
Table 18: Predicted Mean Comments by Content Origin 52
Table 19: Predicted Mean Reshares by Content Origin 53
Table 20: Frequency of engagement depth levels by content origin 54
Table 21: Standardized residuals for engagement depth 54
Table 22: Odds ratios from multinomial logit model (reference = Superficial) 55
Table 23: Interaction effect sizes for AI-only and Mixed groups (ORs for Origin × Exposure) 55

1 Introduction

Generative Artificial Intelligence (GenAI) systems are becoming capable of generating ideas that are considered novel, original, and unique (Guzik et al. 2023). These generated artifacts in the form of text, images, video or even music not only accelerate content creation but, in some contexts, outperform human creators in direct competition (Bauer et al. 2024; Roose 2022). This development has fundamentally changed how creative content is produced and consumed (Atkinson and Barker 2023). In the context of social media, GenAI is used to "create everything from artwork to fully automated social media posts, leading to a surge in AI-generated content" (Park et al. 2024, p. 1). This content sparks significant interest, with Instagram hashtags like #aiart, #aiartcommunity, #aiartwork and #aiartist surpassing 34 million total posts and individual AI influencers with over 2.5 million followers (Alboqami 2023; Instagram 2025).

However, the rapid adoption of GenAI also brings critical concerns that shape how these systems are perceived and integrated into society. These include fears of decreased human creativity (Zhou and Lee 2024), the spread of misinformation (Vasist and Krishnan 2022), and the reluctance to disclose GenAI content (Draxler et al. 2023). In response to these concerns, calls for greater transparency have emerged and policies such as the EU's Article 52 of the European AI Act and the US Executive Order on Safe, Secure, and Trustworthy AI encourage creators to disclose GenAI usage in their work (Bauer et al. 2024).

While these disclosure practices aim to promote trust and accountability, emerging evidence suggests they may have unintended consequences. Recent studies reveal increased negative emotions, perceived threats, and a reduction of perceived quality as responses to GenAI content, especially when when their AI origin is disclosed (Bauer et al. 2024; Gabbiadini et al. 2024). In contrast, Park et al. (2024) report positive user reactions to GenAI content when the origin of the artifact is unknown and highlight the difficulty users face in distinguishing between AI-generated and human-created content.

These effects are particularly relevant in social media contexts, where labeling remains voluntary rather than mandatory, creating a dynamic environment where users' perceptions and reactions to GenAI content can vary widely. While labeling on social media platforms aims to foster transparency and trust, it also introduces a complex interplay between user perceptions, emotions, and engagement with GenAI content. The contrasting findings, negative emotions and reduced perceptions of quality when AI involvement is disclosed versus positive responses when the origin is unknown, highlight a critical gap in our understanding of how transparency influences users' emotional responses.

These diverging outcomes underline the need to explore the emotional and cognitive dynamics at play when users encounter AI-generated content in real-world social media settings. By investigating how content origin affects emotional responses, this study aims to provide actionable insights for designers, practitioners, and policymakers, addressing the question: How do users' emotional responses to explicitly tagged AI-generated content compare to their responses to human-created content on social media?

To address this question, this thesis employs an empirical research approach using quantitative analysis of social media comments. Data from explicitly labeled AI-generated and traditional human-created content are analyzed utilizing advanced Large Language Models (LLMs) to systematically categorize user emotions, sentiment, and engagement behaviors. This method enables a rigorous exploration of how transparency and content origin influence user perceptions and interactions in a real social media environment.

The remainder of this thesis is structured as follows: Chapter 2 provides a comprehensive review of relevant literature, focusing on transparency, emotion, and user engagement with AI-generated and traditional art. Chapter 3 introduces the research model and defines the hypotheses to be tested. Chapter 4 details the research methodology, outlining the approach, data collection, and analysis techniques used to measure emotional reactions and engagement behaviors. Chapter 5 presents empirical results, comparing emotion and engagement metrics between explicitly labeled AI-generated and human-created art. Chapter 6 discusses these findings, interpreting their implications for theory, practice, and policy, while acknowledging study limitations and suggesting areas for future research. Chapter 7 concludes the thesis by summarizing key insights and contributions.

2 Literature Review

2.1 Foundations of Artificial Intelligence in Creative Contexts

In recent years, GenAI has evolved from a niche research interest into a transformative force across creative domains. While early AI systems were characterized by their reliance on rule-based logic or statistical classification, GenAI introduces a paradigm shift: systems that do not merely recognize or categorize input, but generate novel content such as text, images, audio, or video with outputs that are probabilistic, multimodal, and non-deterministic. Although public discourse increasingly uses the term AI synonymously with GenAI, it is important to recognize that GenAI represents a specific subset of artificial intelligence, an advancement nested within the broader AI landscape as depicted in Figure 1 (Banh and Strobel 2023; Sengar et al. 2024).

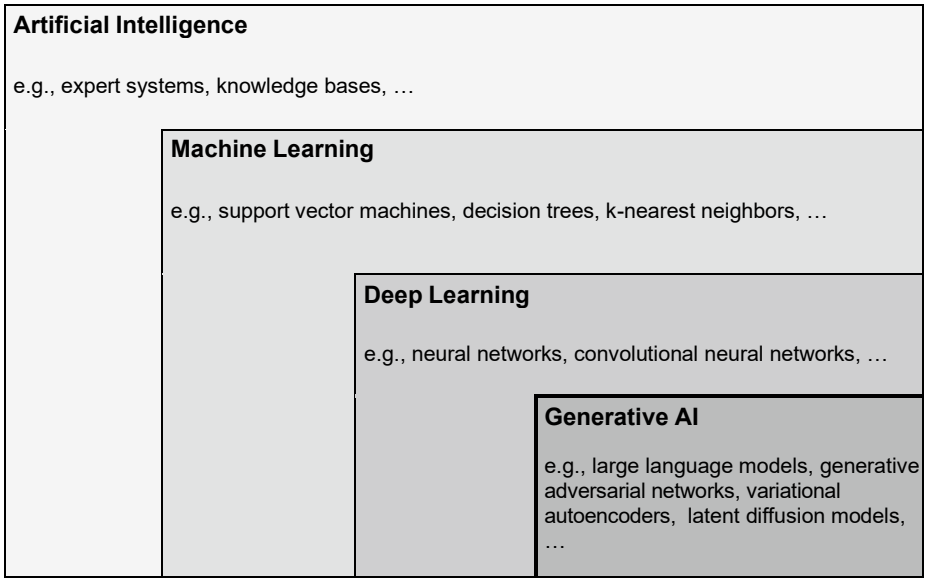

Figure 1: Generative AI and other AI concepts, adapted from Banh and Strobel (2023, p. 2), inspired by Goodfellow et al. (2016, p. 9) and Janiesch et al. (2021, p. 687)

At its core, “Generative AI encompasses artificial intelligence systems with the ability to create text, images, and/ or various forms of media through the utilization of generative models” (Sengar et al. 2024, p. 2). These generative models represent a fundamental departure from traditional discriminative approaches: whereas the latter are designed to classify or predict by learning decision boundaries, the former aim to model the underlying data distribution and generate new, coherent samples from it. This conceptual shift has paved the way for a new class of AI systems capable not only of analyzing information, but of producing creative, context-sensitive outputs.

Among the most prominent realizations of this paradigm are Large Language Models (LLMs), whose development was catalyzed by the introduction of the Transformer architecture (Vaswani et al. 2017). The Transformer’s self-attention mechanism enabled highly effective modeling of long-range dependencies in textual data, setting the stage for generative capabilities at scale. OpenAI’s GPT-3 (Brown et al. 2020) demonstrated the transformative potential of this architecture, achieving unprecedented performance in generating human-like text from simple prompts. Today, LLMs have become the most widely adopted class of GenAI systems, not only due to their versatility in language tasks, but also because they serve as the primary interface through which users interact with generative systems across modalities, leveraging natural language as a universal control layer (Banh and Strobel 2023; Zhou and Lee 2024).

Within this context, critical questions of creativity arise. Creativity has traditionally been defined as the ability to produce outputs that are both novel and appropriate within a given cultural or disciplinary context (Boden 1998). However, as Atkinson and Barker (2023) emphasize, creativity is not a fixed or purely cognitive trait but a socially constructed category, shaped by institutional norms and discursive negotiations about what is considered valuable, original, or even authentically human. The integration of GenAI into creative workflows compels a re-evaluation of how creative agency is attributed and how artistic legitimacy is negotiated in increasingly hybrid human–machine systems.

While early debates on machine creativity - pioneered by Boden (1998) - focused on whether AI could exhibit traits such as originality or intentionality, the current generation of GenAI models has begun to redefine the boundaries of creative labor. Empirical research now shows that these systems can pass established creativity assessments, such as the Torrance Test of Creative Thinking, and generate outputs rated as equally or even more novel than those created by human counterparts (Guzik et al. 2023).

Recent findings complicate this picture. Doshi and Hauser (2023) demonstrate that GenAI-assisted storytelling increases perceived creativity and enjoyment, particularly for users with lower baseline creative ability. However, their study also finds increased output similarity, suggesting that GenAI systems may amplify convergence and limit variation within the creative space. Zhou and Lee (2024) further extend this line of inquiry by analyzing millions of artworks generated using text-to-image systems. Their findings reveal a dual dynamic: while GenAI enhances productivity and peer recognition, it simultaneously leads to declining visual and conceptual novelty at the aggregate level. These trends point to GenAI's role as both a creative equalizer and a driver of aesthetic homogenization.

This transformation has found fertile ground in creative industries. Text-to-image models such as DALL·E 2 and Midjourney are widely used in visual design, marketing, and illustration (Cetinic and She 2021), while generative music systems and video synthesis tools are becoming standard in audiovisual production workflows (Sengar et al. 2024). In practice, GenAI tools are not only augmenting human creativity but in some domains replacing traditional creative processes altogether (Atkinson and Barker 2023; Wu et al. 2021).

What makes this revolution particularly striking is the speed and scale of its adoption. Following the open sourcing of Stable Diffusion in 2021 and the public release of ChatGPT in November 2022, GenAI platforms experienced exponential growth. ChatGPT alone reached over 100 million users in just two months, making it the fastest-growing consumer application in history (Ooi et al. 2025). These developments have triggered a rapid proliferation of GenAI tools and platforms, lowering barriers to entry and enabling non-technical users to engage in complex creative processes via natural language prompts and API-based workflows (Banh and Strobel 2023; Doshi and Hauser 2023).

In parallel, a broader sociotechnical discourse has emerged. On one hand, GenAI is praised for democratizing creativity and increasing productivity. On the other, it raises profound ethical and societal concerns from the erosion of human authorship and artistic originality, to legal disputes over intellectual property and fears of automation-driven displacement in creative industries (Jiang et al. 2022; Sengar et al. 2024). Economic projections estimate that GenAI could boost global productivity by up to 1.5% annually, contributing as much as $7 trillion to global GDP in the coming decade (Ooi et al. 2025). Yet these gains are accompanied by concerns about transparency, accountability, and the long-term cultural value of GenAI content (Atkinson and Barker 2023; Zhou and Lee 2024).

To make sense of this rapidly evolving ecosystem, scholars have proposed layered conceptual frameworks. Banh and Strobel (2023, p. 7) outline a "model-connection-application" stack to describe GenAI infrastructure: foundational models like GPT and Stable Diffusion, linked through APIs and

toolkits, and surfaced through end-user tools such as Jasper, Adobe Firefly, and Midjourney. This modular design has accelerated creative deployment and reshaped participation in content production.

Extending this technical framing, Zhou and Lee (2024, p. 7) propose the notion of “generative synesthesia”, a co-creative paradigm in which humans engage in ideation and prompt refinement while GenAI systems handle execution and variation. This workflow centers human creativity not on executional skill but on conceptual exploration and aesthetic filtering, suggesting a shift in what it means to be creative in a hybrid environment.

Taken together, these developments underscore GenAI’s significance in transforming both the production and reception of creative content. Yet as GenAI-generated artifacts proliferate across digital platforms, new questions arise about how users interpret their origins, attribute authorship, and emotionally engage with such content.

2.2 Transparency and Content Origin on Social Media

As generative AI systems increasingly produce creative content on digital platforms, the question of whether and how to disclose the origin of such content has become a central concern in both regulatory and psychological research. Transparency, while widely seen as a normative good in AI ethics and policy discourse, is a complex and multifaceted construct. It extends beyond the mere provision of information and touches upon deeper questions of user trust, emotional response, and perceived authenticity in mediated interactions.

From a conceptual standpoint, transparency in AI is often framed through relational, systemic, and normative lenses. Larsson and Heintz (2020) distinguish between algorithmic transparency and broader forms of AI transparency, highlighting that meaningful transparency must account for the context in which disclosure takes place, including the expectations and interpretive capacities of the audience. Lund et al. (2025) reinforce this perspective, arguing for risk-based, stakeholder-sensitive transparency mechanisms that are adaptable to varying use cases and domains. Rather than a uniform approach, effective transparency must balance information provision with audience relevance, institutional framing, and legal necessity.

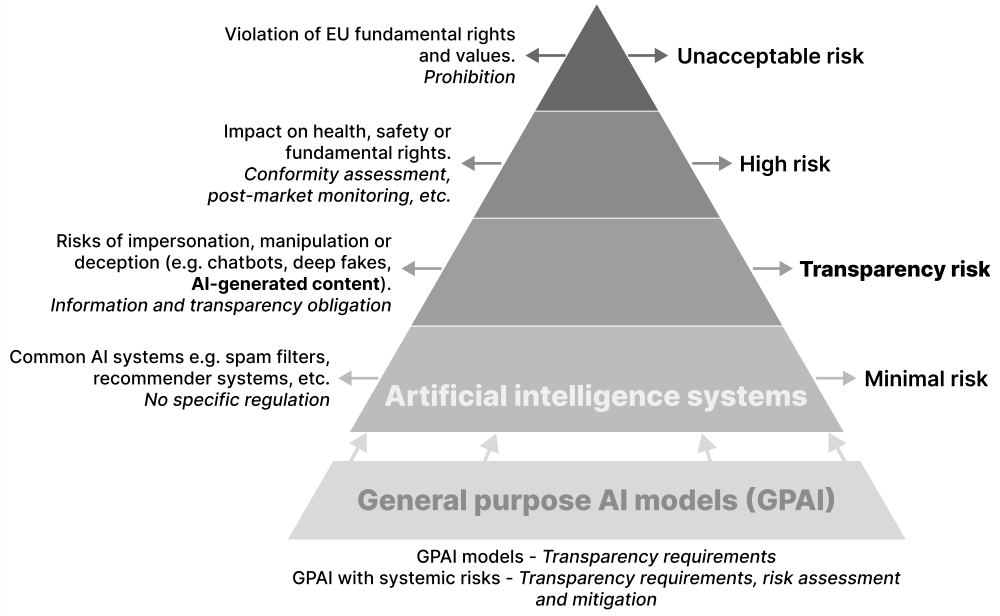

Figure 2: Risk-based pyramid from the EU AI Act, adopted from European Parliament and Madiega (2024, p. 8)

This multidimensional understanding is embedded in the EU AI Act, whose tiered obligations, outlined in Figure 2, place AI-generated or manipulated social-media content in the information-and-transparency tier and, through Article 52, require its artificial origin to be clearly disclosed. As outlined by the European Parliament and Council (2024) the regulation entered into force in mid-2024 and imposes tiered obligations based on the nature and risk of the AI system in question. In parallel, The White House (2023) has introduced Executive Order 14110, which, while less prescriptive, sets a similar trajectory in promoting trustworthy, transparent, and human-centric AI development. These initiatives unfold within an ethics discourse largely shaped by Anglophone institutions, which produce a disproportionate share of global AI-governance documents (Corrêa et al. 2023). Industry responses, such as Meta's evolving AI content labeling policy, illustrate how transparency requirements are operationalized at the platform level. In 2024, Meta introduced a dedicated framework for labeling AI-generated and manipulated media in alignment with global policy developments, including Article 52 of the EU AI Act. According to the transparency documentation available at the time, Meta committed to applying an "AI info" label in two principal scenarios: when AI-generated content could be detected using industry-standard metadata indicators, and when users voluntarily disclosed the use of AI tools to generate or modify their content. Notably, this labeling initially extended beyond fully generated media to include content that had been merely edited or modified using AI tools. As stated in an official update from September 2024, such AI-edited content was still marked with the "AI info" label, though its placement was later adjusted to appear less prominently within the post's menu, rather than directly on the post itself (Meta 2024).

By early 2025, the implementation of this policy has shifted toward a more constrained model. As described in Meta's current Help Center documentation, AI-generated labels are now only applied when content contains embedded industry-standard indicators such as C2PA metadata, an open standard that enables the secure tagging of content origin and modification history within the file itself (C2PA 2024), or has been declared by the user as AI-generated at the point of upload. No mention is made of broader automated detection or labeling outside of these cases (Meta Help Center 2025).

This transition reflects a narrowing in scope from earlier public-facing commitments. While Meta's current approach continues to fulfill minimum disclosure obligations under the EU AI Act, namely indicating AI-generated content when it can be reliably identified through embedded provenance data, it now relies exclusively on externally provided metadata and user declarations. In doing so, the platform has deprioritized broader detection strategies such as visual analysis or platform-internal classification, thereby raising further questions about how transparency standards are interpreted and enacted at scale.

Despite these initiatives, empirical findings complicate the presumed link between transparency and positive user outcomes. Controlled experiments indicate that 'AI-created' labels reduce perceived profundity and monetary worth of art, whereas the same images tagged 'human-created' are rated higher (Bellaiche et al. 2023). Studies by Bauer et al. (2024) and Park et al. (2024) further demonstrate that provenance disclosure can significantly influence user perceptions, frequently in adverse ways. When content is explicitly labeled as AI-generated, users often perceive it as less authentic, less creative, and less emotionally engaging than similar content without such labels. Park et al.'s Instagram-based experiments additionally highlight a measurable engagement gap between labeled and unlabeled AI content, with labeling triggering moral disapproval and reduced trust, even when visual quality remains constant. These findings underscore the emotionally charged nature of transparency in creative contexts.

Gabbiadini et al. (2024) extend this analysis by linking AI disclosure to negative emotional responses, including symbolic threat, anxiety, and resentment. These responses may be shaped by perceptions of

AI intruding into domains traditionally reserved for human expression, such as artistic labor and emotional communication. From a psychosocial standpoint, provenance labels function not merely as informational cues but as framing devices that reposition content from a product of human intention to one of automated generation. In this sense, labeling transforms the interpretive context through which content is received.

These effects align with the broader theoretical insights of Signaling Theory (Spence 1973), which holds that cues such as labels influence judgments of quality and effort. AI disclosures, within this framework, may inadvertently signal reduced authenticity or diminished human contribution, thereby undermining credibility. Media Richness Theory (Daft and Lengel 1986) further suggests that not all modes of disclosure are equally effective. Therefore static labels like “AI-generated” may fail to convey the complexity of generative processes, whereas richer, more layered mechanisms such as interactive model cards or contextual overlays could mitigate misinterpretation and foster more nuanced audience responses.

Lund et al. (2025) add an important architectural perspective, emphasizing the role of continuous, audience-adjusted transparency strategies embedded directly into the design of AI systems. This principle of “transparency by design” represents a shift from reactive to proactive governance, in which disclosure is not a supplementary feature but a core structural component. Notably, standards such as C2PA metadata exemplify this approach: by embedding provenance information such as content origin, editing history, and generative tools used directly into media files, C2PA enables transparency to persist across platforms and contexts (C2PA 2024).

Crucially, while transparency as a concept is intended to support trust and accountability, its implementation may paradoxically undermine these very objectives, particularly in social media environments where creativity, authenticity, and emotional resonance are central to user engagement. Understanding these dynamics requires a closer examination of how users emotionally respond to AI-labeled content, and how emotion itself functions in digitally mediated interactions.

2.3 Theoretical Foundations of Emotion

Emotion, as studied in social contexts such as social media, remains a deeply interdisciplinary concept shaped by diverse theoretical lenses. Mauss and Robinson (2009) define emotion as a multi-component psychological response to personally meaningful events, comprising subjective feeling, physiological arousal, and expressive behavior. This componential model captures the dynamic interplay of cognitive, bodily, and social factors and serves as a foundational definition in the present work. Although originally formulated in broader psychological settings, this definition is especially applicable to digital environments: even where bodily and behavioral cues are obscured, individuals still convey emotion through textual language, making the linguistic dimension central to online emotional expression.

One of the most influential lenses in emotion theory posits that emotions can be divided into a limited number of biologically grounded, discrete categories. According to Ekman (1992), emotions like anger, disgust, fear, happiness, sadness and surprise are universal across cultures, expressed through innate facial configurations, and underpinned by distinct neurobiological circuits. These basic emotions are considered to have evolved to serve adaptive functions, preparing the organism for specific behavioral responses. Plutchik (1980), in a complementary psychoevolutionary model, conceptualizes emotions as fundamental to survival and organizes them into primary dyads, structured by intensity and polarity (e.g., serenity → joy → ecstasy). Both models have played a foundational role in psychological research and continue to shape applied emotion classification efforts.

Contemporary operationalizations of this discrete lens on emotion include the Discrete Emotions Questionnaire (DEQ), developed by Harmon-Jones et al. (2016). The DEQ provides an empirically validated set of eight state emotions: anger, fear, anxiety, sadness, happiness, relaxation, desire, and disgust, designed for self-report studies but grounded in extensive research from affective neuroscience, developmental psychology, and expression analysis.

In more recent work, Cowen and Keltner (2017) extend the discrete approach by empirically mapping 27 distinct emotional categories derived from participants' responses to short video clips. Their findings reveal that emotional experience spans a wider range than suggested by traditional emotion theories. These insights build the foundation for GoEmotions, a large-scale textual emotion taxonomy introduced by Demszky et al. (2020). GoEmotions consists of 27 fine-grained emotion categories (plus a neutral label) annotated from over 58,000 Reddit comments and is grounded in Cowen's semantic space of emotion. Crucially, GoEmotions is not exclusively a linguistic artifact but reflects how emotional categories are meaningfully and reliably expressed in informal digital communication. The GoEmotions taxonomy differs from biological essentialist views by focusing on how emotions are perceived and expressed in everyday textual language. It acknowledges that emotional experiences are often subtle, overlapping and contextually mediated characteristics that are especially pronounced in online interactions. Its design permits multi-label annotation, reflecting the co-occurrence of emotional tones in user-generated content and supporting a more granular interpretation of affective meaning in social media.

From a contrasting lens, theoretical perspectives question the assumption of discrete emotional essences. With Conceptual Act Theory, Barrett (2006) argues that emotions do not arise from fixed biological programs but are constructed through the categorization of core affective states, valence and arousal, using prior conceptual knowledge. Affect, in this context, refers to the basic, pre-conceptual experience of feeling, often described along continuous dimensions such as valence (pleasant–unpleasant) and arousal (low–high intensity), which serve as the foundation for more elaborated emotional experiences. Emotions, in this view, are not directly triggered or perceived, but actively interpreted based on learned social and linguistic frameworks. This lens is especially relevant in digital contexts where users interpret others' expressions not as biological signals but through shared cultural narratives. While Barrett (2006) emphasizes the constructed nature of emotions, dimensional models offer a complementary and measurement-focused perspective. These models propose that emotional experiences are best represented along continuous axes such as valence and arousal. The circumplex model (Russell 1980), maps affective states in a two-dimensional space and has been widely used in affective computing and communication studies. Eerola and Vuoskoski (2011), applying this framework to music-induced emotion, illustrate how dimensional models offer flexibility for interpreting ambiguous or aesthetic experiences, making them particularly useful for analyzing expressive responses to creative content.

Adding a cognitive-relational layer, Appraisal Theory (Smith and Lazarus 1993) proposes that emotion results from an individual's evaluative judgment of a situation's relevance to their goals, values, or well-being. Emotions, according to this theory, are not merely reactions but are shaped by meaning-making processes: whether a stimulus is threatening, controllable, or congruent with expectations determines the resulting emotional experience. This theory complements both dimensional and categorical models by focusing on the interpretive mechanisms that precede emotional expression, an insight critical to understanding how users emotionally respond to AI-generated or human-created content.

Notably, culture affects the appraisal of emotion. Cross-cultural research shows that appraisal and verbal expression of emotion vary systematically with cultural value patterns. Hofstede (2011) presents a framework where low-context, high-individualism cultures (e.g., the Anglophone West) value direct

emotional expression, whereas high-context, collectivist cultures favor harmony-preserving indirectness. Linguistic studies echo this asymmetry: multilingual speakers report greater ease and directness when expressing emotion in English than in languages associated with stricter display rules (Dewaele 2010).

While acknowledging the debate between natural-kind and constructionist views, this work defines emotion in social media as the categorization of affective meaning in text, inferred from user expressions and interpretive contexts. The discrete categorical model GoEmotions serves as the operational framework for the analysis. Its use reflects not a theoretical endorsement but a methodological strategy aligned with the study's aim: to examine how users emotionally respond to AI-labeled versus human-created content in real-world digital environments.

2.4 Emotion in Response to AI and Digital Artifacts

As GenAI becomes increasingly embedded in creative production and digital culture, a growing body of research seeks to understand how users emotionally respond to AI-generated content. These emotional reactions are shaped not only by the content's aesthetic properties but also by users' perceptions of authorship, creative agency, and authenticity. Building on the theoretical foundations of emotion (2.3), and the role of transparency in social media contexts (2.2), this chapter explores how AI-generated artifacts elicit affective responses, including patterns of fascination, discomfort, and symbolic threat.

Building on the early experimental psychology of curiosity, Berlyne (1960) proposed Arousal-Curiosity Theory, which explains how stimulus properties such as novelty, complexity, and ambiguity generate an optimal level of arousal that motivates exploratory behavior. Moderate novelty tends to evoke positive, approach-oriented states, termed epistemic curiosity, whereas very low or very high novelty can produce boredom or aversive tension.

Gabbiadini et al. (2024) provide systematic evidence on emotional reactions to AI-generated content: Across three experimental studies, they expose participants to text, image, and audio materials that were either AI-generated or human-created, controlling for visual and qualitative content characteristics. Participants were asked to rate their emotional responses using pre-defined scales, categorizing them along discrete affective states such as anger, anxiety, and alienation. Significantly higher levels of negative emotional response are reported for AI-generated content compared to human-created analogues, regardless of the media format.

Crucially, these findings parallel core arguments regarding transparency: namely, that transparency, especially when content origin is disclosed, modulates user perception in emotionally charged ways. Gabbiadini et al. (2024) attribute this effect to a form of symbolic threat, wherein users perceive AI-generated content as challenging uniquely human domains such as creativity, emotional expression, and cultural authorship. In this sense, content labeling becomes not only a cognitive cue but also an affective trigger, echoing the function of provenance indicators described in transparency policy frameworks (e.g., EU AI Act, Meta's AI labeling framework). These results provide further evidence that transparency, while normatively encouraged, may yield adverse psychological consequences in emotionally salient domains like art and social communication.

Relatedly, Cheng et al. (2022) investigate emotional responses to AI in professional settings, specifically analyzing how employees react to AI integration in workplace decision-making. Their findings highlight co-existing emotional states, including both relief and anxiety, depending on the perceived purpose and interpersonal framing of the AI system. Emotion is measured via structured surveys targeting self-reported affective states. This mixed-emotion dynamic underscores the appraisal component of emotional response, where emotion arises not from the stimulus per se, but from the perceived

congruence of the stimulus with personal goals, control beliefs, and contextual expectations (Smith and Lazarus 1993). When AI is perceived as a collaborator or assistant, positive affect may dominate; when it is seen as a threat to autonomy or competence, negative affect emerges.

This ambivalence is further explored in Gkinko and Elbanna (2022), who focus on user interactions with conversational AI (chatbots) in emotionally sensitive scenarios. Their analysis, based on qualitative interviews and user diaries, reveals a broader emotional range, including tolerance, hope, and even empathy toward the AI system. Here, users don't rely solely on traditional anthropomorphic cues but respond to the chatbot's perceived communicative effort and contextual sensitivity. The emotional expressions reported map onto both discrete emotions and broader engagement-oriented affective states, demonstrating again the relevance of combining discrete emotion theory with socially constructed meaning frameworks in digital environments.

These findings support a dual-pattern model of emotional response to AI-generated content: one pole oriented toward fascination, hope, and relational empathy, the other toward symbolic threat, anxiety, and alienation. This polarity reflects GenAI's dual role as both a creative enabler and a disruptor. Users are drawn to the novelty, efficiency, and accessibility that GenAI affords, yet simultaneously repelled by the erosion of traditional authorship and emotional authenticity. These emotional patterns are not stable dispositions but fluctuate with context, framing, and the evolving discourse surrounding AI.

This ambivalence is conceptually aligned with the Uncanny Valley Hypothesis (Mori et al. 2012), which suggests that artificial agents that closely resemble but do not perfectly emulate human traits elicit discomfort. While originally developed in robotics, this hypothesis has been extended to digital artifacts. AI-generated art or writing that mimics human creativity may fall into this "valley," provoking adverse reactions precisely because of its near-human quality. In this respect, emotion acts as a boundary marker for human distinctiveness, with discomfort emerging when users sense that the boundary is being crossed.

In parallel, emotional responses to AI content may contribute to more stable cognitive judgments such as algorithm aversion or algorithm appreciation. Negative affective responses, particularly those rooted in symbolic threat or emotional alienation, are likely precursors to distrust, resistance, or withdrawal from AI-mediated interactions. Conversely, positive emotional responses, especially those grounded in empathy or perceived helpfulness, may lay the groundwork for affective trust in AI systems. In sum, the emotional impact of AI-generated content cannot be fully explained by traditional measures of media quality or utility, such as perceived realism or aesthetic value. Instead, these reactions are embedded in users' broader interpretive frameworks, shaped by cultural narratives, transparency cues, and the psychological underpinnings of emotion.

2.5 Algorithm Aversion and Trust in AI Systems

Building on the emotional foundations laid in the previous chapter, this section examines how initial affective reactions to AI-generated content may evolve into more stable cognitive judgments. Central to this transformation is the construct of trust, which functions as a mediating mechanism in the development of long-term attitudes toward AI systems. While emotion often initiates user perception, trust governs the continuity of interaction, shaping whether users engage with, accept, or reject AI-generated content over time. As Gillath et al. (2021) suggest, affective trust, rooted in feelings of interpersonal security can emerge from emotionally positive experiences with AI systems. Conversely, symbolic threat or emotional alienation, often erodes trust, reinforcing disengagement and skepticism (Gabbiadini et al. 2024).

A prominent manifestation of such skepticism is algorithm aversion, first conceptualized by Dietvorst et al. (2014). Algorithm aversion describes the tendency to prefer human over algorithmic judgment, particularly after witnessing algorithmic errors, even when the algorithm outperforms human decision-making in aggregate. This phenomenon stems from a lowered tolerance for machine errors and the belief that algorithms, unlike humans, should perform flawlessly. In domains characterized by subjective interpretation, such as art or creative writing, this aversion is magnified by the perception that algorithms lack intentionality, emotional depth, and contextual sensitivity. As a result, users may acknowledge the efficiency of AI while simultaneously resisting its encroachment into spaces traditionally considered uniquely human.

However, this aversion is not universal. Under specific conditions, users demonstrate what Logg et al. (2019) term algorithm appreciation, an increased preference for algorithmic outputs. Their research reveals that when tasks are perceived as objective or analytically complex, users are more likely to defer to algorithmic decisions. In such contexts, algorithms are seen as more consistent, impartial, and less prone to cognitive bias. This duality underscores that trust in algorithmic systems is not a fixed disposition but a contingent appraisal, sensitive to domain characteristics and task framing.

The formation of trust in AI is also shaped by individual psychological predispositions. Gillath et al. (2021) find that trust in AI correlates with users' attachment styles, highlighting the influence of relational templates on technology interaction. Individuals with secure attachment styles tend to express higher levels of trust in AI, whereas those with anxious or avoidant orientations are more likely to be suspicious or avoidant. This evidence supports the view that trust in AI is not merely a rational judgment of system performance but reflects deeper psychological schemas users bring into the interaction.

Theoretical and empirical work from Bach et al. (2022) provides further granularity, synthesizing findings from human-computer interaction research. Their systematic review identifies three principal antecedents of trust in AI-enabled systems: user characteristics, technical and design features, and socio-ethical considerations. Importantly, they emphasize that trust emerges from the alignment of these factors and is not reducible to any single dimension. While Gillath et al. (2021) foreground individual differences, Bach et al. (2022) extend this perspective by demonstrating how trust is co-constructed between users and systems over time, shaped as much by design as by user psychology.

Lukyanenko et al. (2022) complement this understanding by offering a system-theoretic foundation for trust in AI. Their Foundational Trust Framework conceptualizes trust as an emergent property of interaction between humans and AI systems. Trust, in this view, is not a fixed attribute of a system but a dynamic relational state continuously recalibrated through user experience, system behavior, and broader sociotechnical structures. While Bach et al. (2022) focus on empirical antecedents, Lukyanenko et al. (2022) highlight the ontological status of trust as a process rather than a static outcome. Moreover, they underscore the importance of explainability, perceived control, and institutional embedding in fostering trust. These distinctions reveal a productive complementarity between psychological, interactional, and systems-level analyses of trust.

Together, these perspectives provide a multidimensional account of trust in AI. While aversion may arise from perceived loss of agency or emotional discomfort, trust and appreciation are possible when users are supported by interpretive tools, contextual transparency, and emotionally secure interactional frames. Especially in creative contexts where questions of authorship, meaning, and authenticity are salient, trust functions not only as a judgment of functionality but as a reflection of interpretive alignment between user and system.

In sum, trust in AI systems is not a static quality but a relational achievement. Algorithm aversion and appreciation represent opposing trajectories within a broader spectrum of user engagement. By

understanding the psychological, interpersonal, and systemic dimensions that structure these trajectories, we can better anticipate how users respond to AI-generated content and how those responses evolve into enduring attitudes. This understanding is critical not only for theoretical completeness but also for designing AI systems that are trustworthy, intelligible, and socially acceptable in emotionally salient domains such as digital creativity.

2.6 User Engagement on Social Media Platforms

Understanding user engagement on social media platforms requires a conceptual foundation that accounts for the complex interplay of behavioral, cognitive, and emotional components. Within foundational engagement theory, Brodie et al. (2011) define customer engagement as a psychological state arising from interactive and co-creative experiences with a focal object, such as a brand or platform. This perspective positions engagement not as a single action but as a dynamic process involving cognitive attention, emotional connection, and behavioral response. Hollebeek et al. (2014) builds on this foundation by defining consumer brand engagement as a positively valenced activity manifesting across cognitive, emotional, and behavioral dimensions.

Di Gangi and Wasko (2016) introduce a socio-technical theory of engagement. Their approach highlights engagement as a psychological state shaped through interactions between users and platform-specific features, grounded in individual involvement and the perception of personal meaning. This framing underscores that engagement is not only a response to content or interface design but is also deeply embedded in users' social and interpretive contexts.

Trunfio and Rossi (2021) consolidate these prior perspectives by emphasizing the multidimensional nature of social media engagement. In their systematic review, they argue that engagement in digital environments encompasses not only observable actions such as likes or shares but also latent cognitive and emotional investments, influenced by user-generated content and platform affordances. Their work situates engagement as a dynamic and context-sensitive phenomenon, reflective of both media structures and user interpretation.

A widely applied framework for categorizing user behavior on social media is the COBRA model introduced by Schivinski et al. (2016), which distinguishes three levels of engagement: consumption (passive viewing), contribution (likes, shares, comments), and creation (original content production). This typology aligns with increasing degrees of involvement, from passive reception to active participation. Trunfio and Rossi (2021) emphasize that while cognitive and emotional engagement are central to understanding user experience, behavioral engagement remains the most empirically tractable component. It provides observable, quantifiable indicators that allow researchers to assess user interaction patterns in a scalable way. The authors reference the COBRA framework to underline the importance of contribution-level behaviors, particularly likes and comments as meaningful yet accessible measures of engagement. Both Schivinski et al. (2016) and Trunfio and Rossi (2021) caution that behavioral engagement does not necessarily capture the full depth of user experience. Observable actions may not always reflect the emotional or cognitive intensity underlying them. This distinction is particularly relevant when interpreting engagement with AI-generated content, where measurable responses may diverge from underlying affective reactions.

The motivations that drive social media use further complicate the relationship between emotional response and engagement behavior. According to Uses and Gratifications Theory (UGT), as developed by Katz et al. (1973), individuals actively seek out media to satisfy psychological and social needs, such as entertainment, identity construction, and social interaction. This theory emphasizes user agency and acknowledges that engagement is contingent on the perceived value of the media experience. In

extending UGT to digital contexts, O'Day and Heimberg (2021) highlight how emotional states, particularly loneliness and social anxiety, shape social media behavior. Their findings suggest that users experiencing these affective conditions may turn to social media as a compensatory mechanism, though such use does not always result in deeper or more constructive engagement. Users may interact frequently, but superficially, particularly when the content appears emotionally incongruent or unfamiliar.

This perspective complicates assumptions that emotionally evocative content necessarily leads to heightened engagement. As Verduyn et al. (2017) note in their review of social network site usage, active forms of engagement such as commenting or posting are associated with positive psychosocial outcomes, whereas passive use such as scrolling can exacerbate negative feelings through social comparison and emotional detachment. Trunfio and Rossi (2021) support this differentiation, observing that although passive behaviors dominate usage patterns, active engagement is more strongly associated with meaningful psychological investment.

These distinctions are especially salient when considering platform-specific norms. On image-centric platforms like Instagram, users are exposed to highly curated, visually expressive content. Pittman and Reich (2016) argue that these affordances foster parasocial interaction and emotional immediacy, but also heighten the pressure for aesthetic and social conformity. Familiarity with content types adds yet another dimension to user behavior. Di Gangi and Wasko (2016) propose that engagement is influenced not only by individual dispositions but also by system-level features such as content personalization and the presence of a critical mass. Users regularly exposed to both AI-generated and human-created content may develop differentiated engagement patterns, adjusting their behaviors based on expectations and perceived authenticity. This familiarity may lead to more selective or reserved interaction with AI content, reflecting evaluative processes shaped by prior exposure and interpretive schema.

In summary, user engagement in social media environments represents a multi-layered construct that cannot be reduced to surface-level interaction metrics. Theories of engagement, typologies of behavior, and motivation-based models collectively illustrate that what users do on social platforms is deeply shaped by how they feel, what they seek, and the structures within which they operate.

2.7 Emotion Detection in Text: Capabilities and Methods

Emotion expression is integral to human communication, emerging through verbal, non-verbal, and textual modalities, and reflecting the multidimensional structure of affective processes as described by Mauss and Robinson (2009). In digital contexts, particularly on social media, text has become the predominant carrier of emotional signals. As such, understanding how emotions are embedded in language has become a central challenge for natural language processing (NLP), especially in domains requiring sensitivity to psychological and social meaning. Emotion detection in text has thus evolved into a critical capability for a range of applications, including adaptive recommender systems, content moderation, sentiment-aware chatbots, and mental health monitoring tools (D'Andrea et al. 2015; Nandwani and Verma 2021).

Unlike sentiment analysis, which typically assesses general evaluative tone (positive, negative, neutral), emotion detection aims to classify discrete emotional states such as joy, fear, or admiration based on linguistic signals. This distinction requires greater semantic granularity and is often informed by taxonomies rooted in psychological models (D'Andrea et al. 2015; Demszky et al. 2020). In text-based environments, where non-verbal and physiological indicators are absent, emotion detection must rely exclusively on language structure and context, making methodological precision especially critical.

Emotion detection, or emotion recognition in NLP, is generally defined as the process of identifying and classifying emotional states conveyed in written language (Nandwani and Verma 2021). Nandwani and Verma (2021) characterize it as a subfield of affective computing, aiming to recognize affective states within textual data. D'Andrea et al. (2015) reinforce this position, distinguishing emotion detection from sentiment analysis and highlighting its higher granularity and interpretive demands. The GoEmotions dataset by Demszky et al. (2020) represents a notable empirical effort in this area, comprising over 58,000 English Reddit comments annotated with 27 emotion categories plus a neutral label, offering a benchmark for fine-grained emotion classification.

Methodologically, three principal approaches have emerged. Lexicon-based models associate predefined word lists with specific emotional categories. These methods, including tools like WordNet-Affect, NRC, and LIWC, offer high interpretability but often struggle with contextual ambiguity and domain transferability (D'Andrea et al. 2015; Nandwani and Verma 2021). Machine learning methods, including Naive Bayes, Support Vector Machines, and Random Forests, have been widely used, typically relying on supervised learning from annotated corpora. These approaches outperform lexicon-based models in adaptability but require extensive feature engineering and manually labeled training data (Bota et al. 2019).

Recent developments have shifted toward deep learning and large language models. Transformer-based architectures, such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), have demonstrated the capacity to encode nuanced contextual dependencies, allowing for improved emotion classification performance (Devlin et al. 2019; Radford et al. 2018). Demszky et al. (2020) report that a fine-tuned BERT model achieved a macro-averaged F1 score of 0.46 on the GoEmotions dataset, outperforming traditional classifiers. These advances display the broader paradigm shift from handcrafted rules and fixed features to models that learn representational structures from large-scale data.

Comparative analyses of LLMs have further sharpened our understanding of their capabilities. Boitel et al. (2024) assess the performance of GPT-3 and DeBERTa v3 on emotion recognition, noting that BERT-based models typically retain superior accuracy due to their task-specific training and domain adaptation mechanisms. However, they also highlight the versatility and generalizability of GPT-based systems, which excel in zero-shot and few-shot scenarios, albeit with some trade-offs in precision. Kocoń et al. (2023) corroborate this perspective by evaluating ChatGPT and GPT-4 across 25 NLP tasks, including emotion detection. Their findings show that while ChatGPT performs adequately, it lags behind state-of-the-art (SOTA) systems by an average margin of approximately 25% in F1 score on emotion tasks, underscoring its limitations in fine-grained affective analysis.

Language and cultural specificity pose additional challenges. The majority of benchmark datasets, including GoEmotions, are in English, which limits cross-linguistic applicability. As Niu et al. (2024) note, most emotion classification systems are trained and evaluated on monolingual corpora, often ignoring cultural variations in emotion expression. This issue is compounded in social media environments, where users frequently code-switch or employ hybrid linguistic forms. Consequently, emotion recognition models trained exclusively on English data may yield lower accuracy on multilingual content or culturally distinct affective expressions.

The role of emojis in text-based emotion detection introduces another layer of complexity. Jahan et al. (2024) highlight the ambiguity of emoji use, demonstrating that interpretation varies significantly across platforms and user demographics. While emojis can signal affective intent, their decoding is context-dependent and often misaligned with intended meaning. These findings caution against assuming a one-to-one mapping between emoji presence and emotional category, particularly in automated classification systems.

Taken together, these findings offer a nuanced understanding of the current landscape in text-based emotion detection. They underscore the strengths of deep learning models, particularly transformer-based architectures, in handling complex affective inference, while also pointing to persistent limitations related to cultural variability, multimodal ambiguity, and contextual interpretation. Building on these insights, the present study implements a two-tier pipeline, leveraging ChatGPT for multilingual, emoji-aware weak labelling and a GoEmotions-fine-tuned BERT for the English subset, the full rationale of which is outlined in the forthcoming methodology chapter.

2.8 Research Gap and Thesis Positioning

The preceding review has traced a layered and interdisciplinary discourse around generative AI, transparency, emotion, and user engagement in digital contexts. These strands converge on three core insights. First, GenAI has become a widely adopted tool in creative domains, especially on image-centric platforms such as Instagram, where it increasingly contributes to the production and circulation of digital art (Alboqami 2023; Park et al. 2024). Second, the labeling or disclosure of AI-generated content, while intended to foster transparency, has been shown to influence user perception, often eliciting negative emotional responses and reducing trust or engagement (Bauer et al. 2024; Gabbiadini et al. 2024). Third, advances in natural language processing, particularly through transformer-based architectures, have enabled scalable emotion detection in user-generated content, making it feasible to analyze affective responses to digital media at scale (Demszky et al. 2020; Devlin et al. 2019).

Despite the conceptual and technical progress outlined above, prior empirical research in this area remains limited in scope. Most existing studies rely on controlled experimental designs using small participant samples and curated stimuli (e.g. Bauer et al. (2024), Gabbiadini et al. (2024), Park et al. (2024)). These approaches are effective for isolating causal mechanisms but are inherently constrained in their ability to reflect the complexity and heterogeneity of real-world social media interactions. Specifically, little is known about how users engage with and emotionally respond to AI-generated versus human-generated creative content within authentic, multilingual, and contextually variable environments such as Instagram.

This thesis addresses this empirical and methodological gap by analyzing a corpus of 64,000 user comments on art-related Instagram posts, comparing interactions with content tagged as AI-generated and human-authored. Drawing on recent developments in natural language processing, the study employs state-of-the-art classification strategies to infer emotion and engagement from textual data. Rather than relying on a single approach, it combines the flexibility of general-purpose language models with the specificity of fine-tuned emotion classifiers, enabling the capture of both broad multilingual patterns and high-resolution affective signals in English-language content. In doing so, the study extends the current literature by providing observational evidence from an organic, large-scale social media setting, complementing existing experimental findings and expanding their external validity.

This thesis is situated at the intersection of generative AI, affective computing, and information systems (IS) research. By integrating theoretical constructs such as transparency, discrete emotion theory, and engagement typologies into a real-world analytical framework, it seeks to advance both the empirical understanding of user response to AI-generated content and the methodological toolkit available for studying affective behavior in digital environments. Through this dual focus, the work contributes to ongoing discussions about the role of generative technologies in public discourse, artistic production, and platform governance.

3 Research Model

The conceptual framework of this study is grounded in prior literature on GenAI, emotion theory, transparency, and digital user engagement. It seeks to examine how Content Origin, defined as whether content is AI-generated or traditionally human-created, influences users' emotional and behavioral responses on social media. Four direct effects are hypothesized: on discrete emotion profiles (H1), sentiment valence (H2), behavioral engagement (H3), and engagement depth (H4). Additionally, the model incorporates User Exposure Pattern as a moderating variable (H5), assessing whether users who engage with both content types display differentiated emotional responses. Transparency, operationalized as the disclosure of AI origin, is included as a control variable, conceptually relevant for interpretive context but not subject to hypothesis testing.

3.1 Content Origin and Affective Response (H1, H2)

Emerging empirical evidence suggests that the origin of creative content significantly shapes user emotion. AI-generated works, even when aesthetically comparable to human-authored counterparts, tend to evoke more negative discrete emotions such as anxiety, resentment, or alienation (Gabbiadini et al. 2024; Park et al. 2024). This phenomenon is rooted in symbolic threat perception, whereby AI intrudes into traditionally human domains, such as artistic authorship or emotional expression. From the lens of discrete emotion theory (Ekman 1992; Harmon-Jones et al. 2016), this implies shifts in the distribution of categorical affective responses. Hence:

H1: Content Origin (AI-generated vs. human-created) significantly alters the distribution of discrete emotional expressions in user comments.

Further, dimensional models of affect suggest that emotional experience also varies along valence and arousal axes (Russell 1980). Prior research shows that AI-origin cues lead to less positive and more ambivalent sentiment evaluations, even when controlling for visual quality (Bellaiche et al. 2023; Gabbiadini et al. 2024). These effects reflect not only emotional nuance but interpretive reappraisal of the content's authenticity and meaning (Smith and Lazarus 1993). Therefore:

H2: Content Origin influences the valence of user sentiment, with AI-generated content eliciting more negative sentiment than human-created content.

3.2 Content Origin and Engagement Behavior (H3, H4)

Behavioral engagement, a core dimension in social media research, is defined here via observable metrics such as likes, comments, and child comments (Schivinski et al. 2016; Trunfio and Rossi 2021). While emotionally evocative content typically drives higher engagement (Brodie et al. 2011), prior work cautions that negative affect linked to AI disclosure may suppress interaction due to moral disapproval or reduced perceived authenticity (Gabbiadini et al. 2024; Park et al. 2024). Consequently:

H3: Content Origin affects behavioral engagement, with AI-generated content eliciting significantly different levels of likes, comments, and replies than human-created content.

Beyond frequency, engagement depth, defined for this research project as the cognitive and emotional elaboration evident in user responses, captures more subtle forms of involvement. According to theories of co-creative value and emotional immediacy, perceived authenticity enhances the meaningfulness of

user expression (Di Gangi and Wasko 2016; Trunfio and Rossi 2021). As prior studies show, AI-generated content reduces the perceived emotional resonance and interpretability of creative work (Bauer et al. 2024), potentially leading to shallower, less reflective commentary.

H4: Content Origin affects the depth of user engagement, with AI-generated content eliciting less engagement depth than human-created content.

3.3 Moderating Role of User Exposure Pattern (H5)

A nuanced dimension of the model involves the role of prior user exposure to both AI-generated and human-created content in shaping emotional responses. Literature suggests two contrasting interpretations. On one hand, repeated exposure to both content types may reduce emotional contrast by diminishing symbolic threat and fostering familiarity (Di Gangi and Wasko 2016; Gabbiadini et al. 2024). On the other hand, broader exposure may enhance users' interpretive frameworks, enabling more nuanced evaluations and heightened sensitivity to origin-related cues. A view supported by constructivist emotion theory and appraisal theory, which stress the role of conceptual knowledge and contextual meaning-making in affective appraisal (Barrett 2006; Smith and Lazarus 1993). While the present study does not explicitly predict the direction of this moderating effect, it accounts for the potential interpretive complexity that user familiarity introduces into emotional evaluations. Accordingly, the following hypothesis is proposed:

H5: User Exposure Pattern moderates the relationship between Content Origin and emotional response, such that users exposed to both AI and human-created content exhibit more differentiated emotional profiles than users exposed to a single content type.

3.4 Transparency as a Control Factor

Transparency, particularly the disclosure of content provenance, has been shown to intensify affective and interpretive reactions. Empirical studies demonstrate that labeling content as "AI-generated" significantly reduces perceived authenticity, creative value, and emotional resonance, while simultaneously increasing skepticism and moral disapproval (Bauer et al. 2024; Bellaiche et al. 2023). Such provenance labels function as cognitive and emotional frames, shaping the audience's appraisal of creative artifacts by repositioning them within the interpretive space of artificial production. In the present study, Transparency is not treated as an independent predictor in the main analysis but is included as a control variable. This methodological decision reflects the construct's conceptual relevance, as emphasized by recent regulatory frameworks such as the EU AI Act (European Parliament and Council 2024), without detracting from the primary focus on Content Origin. Given its theoretical and practical significance, Transparency is formalized as a supporting hypothesis, acknowledging its conceptual presence within the research model. Although not empirically tested in this study, it informs the interpretation of findings and serves as a foundation for future research directions.

HS1: Transparency influences emotional and engagement outcomes, with disclosed AI-generated content eliciting less positive emotional responses compared to non-disclosed content.

3.5 Conceptual Research Model

Figure 3 visualizes the conceptual structure of the research model, highlighting the independent variable (Content Origin), dependent variables (Emotion, Sentiment, Engagement, Engagement Depth), moderator (User Exposure Pattern), and control variable (Transparency). Solid arrows represent the primary hypothesized relationships tested in this study, with the moderating role of exposure applied specifically to the relationship between Content Origin and emotion. The dashed arrow indicates the supporting hypothesis (HS1), representing a theorized but untested relationship between Transparency and user responses.

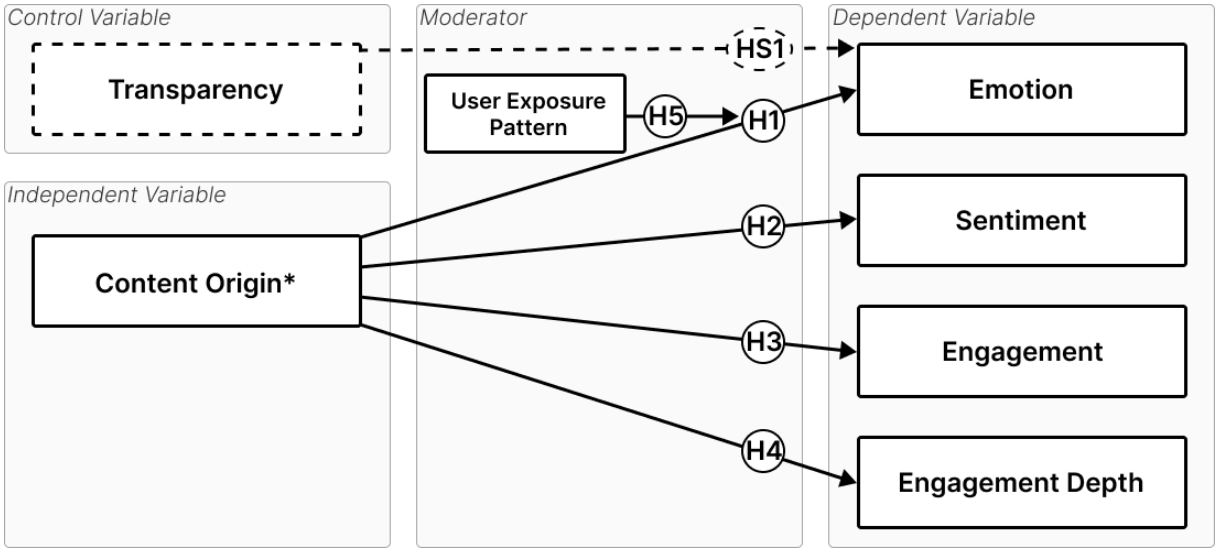

Figure 3: Research Model. *Content Origin is coded only for posts whose artificial origin is transparently disclosed.

Table 1 provides an overview of the five main hypotheses and the supporting hypothesis derived from the model.

Hypothesis	*Statement*
H1	Content Origin (AI-generated vs. human-created) significantly alters the distribution of discrete emotional expressions in user comments.
H2	Content Origin influences the valence of user sentiment, with AI-generated content eliciting more negative sentiment than human-created content.
H3	Content Origin affects behavioral engagement, with AI-generated content eliciting significantly different levels of likes, comments, and replies than human-created content.
H4	Content Origin affects the depth of user engagement, with AI-generated content eliciting less engagement depth than human-created content.
H5	User Exposure Pattern moderates the relationship between Content Origin and emotional response, such that users exposed to both AI and human-created content exhibit more differentiated emotional profiles than users exposed to a single content type.
HS1	Transparency influences emotional and engagement outcomes, with disclosed AI-generated content eliciting less positive emotional responses compared to non-disclosed content.

Table 1: Hypotheses Overview

The research model and hypotheses outlined above provide a structured framework for analyzing user responses to AI-generated and human-created content on social media. The following chapter details the methodological approach used to empirically test these relationships.

4 Research Methodology

This chapter outlines the methodological framework used to investigate how users emotionally and behaviorally respond to AI-generated versus human-created content on Instagram. It specifies the research design, classification pipeline, and modeling strategy employed to test the study's five theory-driven hypotheses. Emphasis is placed on computational techniques, naturalistic data collection, and classifier-assisted annotation procedures that support the empirical analysis of large-scale user-generated content. The approach combines NLP-based labeling, user segmentation, and statistical modeling within a reproducible, ethics-aligned architecture. Methodological decisions are grounded in established practices from affective computing, information systems research, and computational social science.

4.1 Research Design

This thesis adopts a naturalistic, large-scale computational content analysis of user-generated Instagram data, situated at the intersection of affective computing, transparency studies, and IS research on user engagement. The methodological approach combines principles of observational digital research with recent advances in NLP to examine emotional and behavioral responses to AI-generated content in real-world social media environments. While the study is exploratory in its subject matter, addressing the underexplored domain of user emotion in response to disclosed generative AI content, it applies theory-driven hypotheses and structured modeling procedures to test these relationships empirically.

The design is characterized as non-interventional and confirmatory (Creswell 2014), with pre-defined hypotheses derived from established literature in affective science, algorithm aversion, transparency, and social media engagement. Although preliminary qualitative work on a subset of the data, particularly during the design of the engagement depth classification, does inform certain measurement strategies, the study does not primarily pursue inductive theorizing. Rather, it employs these computational classifications to operationalize latent constructs and test previously theorized relationships.

At its core, the research builds on the methodological tradition of computational content analysis, which Grimmer and Stewart (2013) characterize as the application of automated classification methods to analyze unstructured text at scale. This approach enables the detection of meaningful linguistic and emotional patterns across large corpora without direct human coding. In line with their interpretation, the study does not aim to replace human interpretation but to extend its reach, transforming the manual work of labeling thousands of data points into a replicable, model-based inference process. Classification tasks for emotion, sentiment, and engagement depth are executed using LLMs, specifically GPT and BERT variants, with interpretive scaffolding provided by existing emotion taxonomy.

The decision to focus on Instagram content aligns with the principles of naturalistic observational research, widely adopted in computational social science (CSS). Lazer et al. (2020) emphasize the unique strengths of unobtrusively studying human behavior in native digital environments where interactions are self-motivated, contextually embedded, and shaped by real-time social and platform dynamics. By analyzing authentic user comments on publicly available posts tagged as #aiart or #traditionalart, the study captures user reactions as they occur organically, outside the laboratory or experimental prompt setting. The research process follows a structured pipeline, comprising: data collection and cleaning, automated classification of affective and engagement signals, user-level segmentation based on exposure patterns and hypothesis testing through statistical modeling.

Although computational research designs offer notable advantages in scalability and ecological validity, they also entail specific challenges, such as noise in language models, biases in platform dynamics, and the absence of ground truth labels. As Lazer et al. (2020) caution, social media data may be shaped by hidden curation mechanisms, shifting platform norms, and variable metadata fidelity. These limitations are acknowledged and addressed where applicable. A comprehensive account of methodological risks and mitigations is provided in Section 4.12.

Overall, this research design supports the study's goal of quantifying emotional and engagement-related reactions to AI-generated content in a platform-native context. It reflects a pragmatic commitment to leveraging computational tools emphasizing practical, context-sensitive strategies for addressing complex research problems (Creswell 2014), while maintaining awareness of interpretive complexity, ethical responsibility, and the socio-technical fabric in which digital interactions unfold.

4.2 Philosophical Positioning

This study adopts a pragmatic epistemology, emphasizing methodological utility and context-sensitive insight over rigid adherence to any single worldview (Creswell 2014). Rather than subscribing to a fixed paradigm, the approach reflects a pluralistic stance that draws on both quantitative and qualitative techniques to best address the research problem. This orientation supports the integration of computational tools such as transformer-based language models, chosen for their practical efficacy in analyzing complex, informal user-generated data.

The design also reflects elements of post-positivist reasoning, particularly in its use of theory-informed hypotheses and statistical modeling. As Creswell (2014) outlines, post-positivism assumes an objective reality exists, but acknowledges that it can only be approximated through probabilistic inference and fallible measurement. In this study, statistical models are used not to assert definitive claims, but to estimate structured relationships between language-based indicators and conceptual categories such as emotional tone or engagement depth.

Ontologically, the study adopts a constructivist view of emotion, treating it as a context-dependent phenomenon expressed and interpreted through language. This contrasts with essentialist theories that define emotion as a set of biologically fixed states (Ekman 1992; Plutchik 1980). Instead, the analysis draws on Conceptual Act Theory (Barrett 2006), which sees emotion as emerging through the conceptual interpretation of core affective states, shaped by linguistic and social context. From this perspective, large language models are used not to detect internal states, but to identify recurring patterns of expressed affect in public discourse. This aligns with the view that emotion, in digital environments, is not directly observable but constructed through interaction, narrative, and expression.

Taken together, this philosophical positioning supports the study's flexible but theory-aligned use of computational methods to investigate affective phenomena in digital environments.

4.3 Methodological Justification in Information Systems Research

This study is situated within the computational strand of Information Systems Research (ISR), where large-scale digital trace data and algorithmic methods are increasingly employed to examine socio-technical phenomena. While traditionally ISR is dominated by surveys, case studies, and mathematical modeling, recent research reveals a marked rise in data mining, machine learning, and text analytics approaches (Mazaheri et al. 2020).

Rank	*Methodology*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*
1	Survey	49	53	69	45	56	47
2	Mathematical modeling	55	55	47	50	61	72
3	Case Study	46	47	31	33	25	32
…							
14	Machine learning, Data mining, …	2	3	0	5	10	21
15	Content analysis	2	3	5	5	2	2
16	Social Network Analysis	1	2	0	4	1	3

Table 2: Methodology trends over the years 2013-2018, adapted from Mazaheri et al. (2020, p. 12)

Mazaheri et al. (2020) document this shift, showing that although such methods still represent a small share of ISR overall, their annual usage has grown rapidly in top-tier journals, especially since 2015 (see Table 2). This reflects a broader disciplinary openness toward methodological innovation, particularly for research engaging with online behavior and algorithmic systems.

The present work aligns with what Recker (2021) describes as computational research: methodologically rigorous studies that use algorithmic reasoning and scalable inference to investigate complex digital environments. By combining established classification techniques with novel application in the context of AI-labeled social media content, this thesis extends ISR's methodological repertoire without constituting a design science study, as the developed pipeline is not evaluated as a research artifact but serves as a means for empirical investigation. It addresses theory-driven research questions through computational modeling of user-generated data, an approach increasingly recognized as legitimate and valuable in ISR.

This methodological orientation supports ISR's broader goals of understanding human–technology interaction, sociotechnical dynamics, and digitally mediated behavior. It reflects what Recker (2021) frames as a commitment to methodological alignment and rigor, selecting approaches that fit the phenomenon under investigation and enable credible knowledge generation in socio-technical systems.

4.4 Data Source and Collection Procedure

This study employs a minimally invasive and ethically aligned strategy to collect publicly available Instagram content for large-scale computational analysis. The objective is to gather naturalistic, user-generated comments on AI-generated and traditional artworks using scalable methods that preserve privacy and comply with data protection standards. The end-to-end collection procedure is summarized in Figure 4, which outlines the five integrated phases.

Figure 4: Data Collection and Storage Pipeline, *including Anonymization of UserID

Between March 8 and April 8, 2025, data was retrieved using a publicly accessible service API hosted on RapidAPI (https://rapidapi.com/). This tool is selected for its reliability and ability to retrieve public content without requiring authentication or user interaction. As Instagram's official Graph API does not support academic access to post- or comment-level data, web scraping remains a legitimate alternative for scholarly research involving public social media content (Jünger 2023; Peters et al. 2023).

The content selection process targets 'Top' posts sorted by popularity under the hashtags #aiart and #traditionalart, encompassing a variety of post formats including single images, videos, and carousels. Each day, multiple pages of top-ranked posts are retrieved, and their unique identifiers (postID) are

stored in a JSON file. For each postID, the first page of user comments is extracted, with pinned comments excluded to reduce selection bias.

During the retrieval of comments, user identifiers are anonymized using a salted one-way SHA-256 hashing procedure. This design choice aligns with General Data Protection Regulation (GDPR)-compliant privacy strategies, ensuring that no direct or indirect identifiers are stored (Saltarella et al. 2021). Consequently, usernames, profile links, or other personal metadata are never retained. The resulting dataset is stored on encrypted, access-controlled infrastructure and is manually reviewed to confirm integrity and privacy compliance.

Following collection and anonymization, data cleaning removes duplicate entries, particularly those arising from repeated daily crawls. To further comply with the GDPR's storage limitation principle (European Union 2016, Art. 5(1)(e)), raw comment texts are retained in pseudonymised form for up to twelve months, strictly for purposes of replication and audit. After this period, the corpus will be securely deleted. Although the comments analysed are publicly accessible, they are not authored with academic reuse in mind. The study therefore upholds contextual integrity by limiting retention, avoiding direct quotation, and ensuring all data is stored securely and access-controlled (Jünger 2023; Nissenbaum 2010).

The following variables are collected to support the modeling of emotional expression, engagement behavior, and platform interaction patterns:

Post-level Fields		***Comment-level Fields***	
Field	**Description**	**Field**	**Description**
postID	Unique identifier of the post	*postID*	Unique identifier of the associated post
likeCount	Total number of likes	*commentID*	Unique identifier of the comment
commentCount	Total number of comments	*userID*	Hashed userID
reshareCount	Total number of reshares	*text*	Raw comment text
hashtag	Hashtag used to identify content origin	*likeCount*	Number of likes received
		replyCount	Number of replies to the comment
		isPinned	Boolean indicating comment was pinned
		hashtag	Hashtag identifying the associated post

Table 3: Raw Data Fields Structure

While Instagram's Terms of Use generally prohibit scraping through automated means, scholars such as Peters et al. (2023) and Jünger (2023) emphasize the interpretive ambiguity surrounding public data usage in research. The absence of a suitable academic API, combined with the exclusive use of non-authenticated, publicly visible data, supports a research justification based on public interest. Ethical safeguards including the exclusion of private content, proactive anonymization, and timely deletion of raw text serve to mitigate risks and uphold platform trust.

In this study, technical constraints are not treated solely as obstacles but as reflections of Instagram's sociotechnical governance. Rate limiting, content volatility, and undocumented data structures are understood as part of the platform's epistemic boundaries (Jünger 2023, p. 429) and are explicitly acknowledged in the design and interpretation of the dataset.

In sum, this chapter outlines a data collection strategy that combines scientific rigor, privacy protection, and methodological transparency. A comprehensive reflection on broader ethical implications is provided in Chapter 4.13.

4.5 Emotion Classification Pipeline

This study employs a dual-model emotion classification pipeline designed to balance coverage, interpretive richness, and methodological consistency in the extraction of emotional signals from Instagram comments. Building on the conceptual foundations outlined in Chapter 2.7, the classification pipeline serves as the core operational mechanism for testing Hypothesis 1 (H1), which investigates how the origin of visual content (AI-generated vs. human-created) shapes the distribution of discrete user emotions in social media discourse. Emotion detection is not merely a descriptive tool in this context, but a variable-defining mechanism essential for the study's core statistical analyses.

The pipeline is structured around two complementary NLP models: a GPT based model, used for large-scale weak labeling across all comments, and a fine-tuned BERT model, applied selectively to the English-language subset for higher accuracy. The GPT model processes the entire dataset and performs multiple classification tasks in a single pass: it assigns a primary emotion label from the GoEmotions taxonomy (Demszky et al. 2020), infers sentiment polarity, estimates engagement depth, identifies language, and flags sarcasm. These multiple annotation layers serve not only the emotion-related hypothesis (H1), but also subsequent analyses involving user tone, response complexity, and classification robustness.

Following the GPT-based annotation, all comments identified as written in English are routed to a secondary classification phase, in which a BERT model fine-tuned on the GoEmotions dataset is applied. This step generates an additional BERT-based emotion prediction with its own associated confidence score. This parallel labeling enables later comparison between models and facilitates robustness checks on classifier agreement. The rationale for including BERT lies in its demonstrated superiority in task-specific emotion recognition when fine-tuned on structured corpora (Boitel et al. 2024; Demszky et al. 2020; Devlin et al. 2019). Although ChatGPT provides broader multilingual and zero-shot capabilities (Kocoń et al. 2023), BERT's domain adaptation offers stronger reliability on the English subset, especially for emotion classification rooted in discrete taxonomies. This architecture allows the study to harness GPT's multilingual generalizability while maximizing accuracy where benchmarked models exist.

To accommodate downstream analysis and enable robustness checks, the source and confidence of each emotion label are retained as part of the comment's metadata. This practice supports later segmentation and ensures transparency in classifier attribution. The harmonization step also addresses edge cases, such as language misclassification or inconsistent emoji interpretation, by creating a unified schema that aligns all outputs to a single taxonomic standard: the 28-label GoEmotions taxonomy.

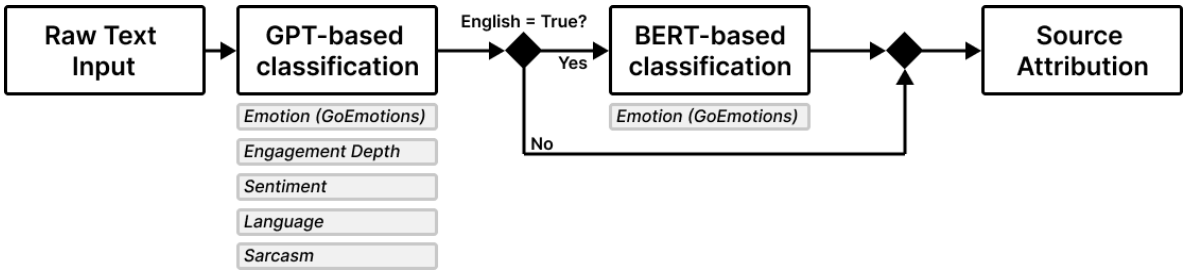

Figure 5: Text Classification Pipeline

A visual overview of the classification pipeline is presented in Figure 5. It depicts the sequential routing from raw Instagram comments to final harmonized emotion labels, showing the full set of annotations produced by the ChatGPT stage, including emotion, engagement depth, sentiment, language and sarcasm, and the refinement step introduced by the BERT classifier on English content. This modular

setup is crucial for managing the inherent challenges of social media data: multilinguality, emoji use, code-switching, and informal expression styles (Jahan et al. 2024; Niu et al. 2024).

Overall, this dual-stage pipeline reflects a methodological compromise between coverage and precision. It acknowledges that while no model alone can fully resolve the ambiguity and contextual complexity of user-generated content, a layered, language-aware architecture provides the best available approximation. The classification output feeds directly into both descriptive statistics and inferential testing, anchoring the study's engagement with affective discourse in reproducible, well-documented NLP procedures.

4.5.1 GoEmotions Taxonomy

The classification of emotional expression in this study is based on the GoEmotions taxonomy, a fine-grained and empirically derived annotation schema introduced by Demszky et al. (2020). Developed to overcome the limitations of sentiment analysis and constrained categorical taxonomies, GoEmotions comprises 27 discrete emotion labels and a neutral class, constructed through manual annotation of 58,000 English Reddit comments. Its design draws on and extends a semantic framework for emotion, which maps human affective experience along a spectrum of nuanced and frequently co-occurring categories (Cowen and Keltner 2017). In contrast to psychologically reductionist approaches focused on biologically hardwired basic emotions (Ekman 1992), the GoEmotions taxonomy reflects a broader, linguistically grounded conception of affective expression in naturalistic digital communication.

The taxonomy includes both evolutionarily foundational emotions such as joy, anger, and fear and a range of cognitively mediated and socially embedded states such as admiration, disappointment, and realization. This diversity reflects the contextual richness of emotion expression on social media platforms and aligns with constructionist perspectives on emotion as interpretively assembled rather than biologically fixed (Barrett 2006). The inclusion of these categories enables more granular analysis of emotional discourse and supports the study's goal of evaluating emotion in response to AI-generated versus human-created content in informal, user-generated texts.

Structurally, the GoEmotions taxonomy retains conceptual compatibility with both discrete and dimensional models of emotion. While it offers a set of distinct emotion categories consistent with discrete approaches (Cowen and Keltner 2017; Ekman 1992), its organization also permits mapping onto the valence dimension of dimensional theories (Russell 1980). Demszky et al. (2020) group the 27 emotion labels into positive, negative, and ambiguous categories based on their affective orientation, allowing for sentiment-level aggregation and interpretive linkage with broader polarity-based measures.

Beyond valence-based classification, the hierarchical structure of the taxonomy reveals semantic proximities among categories. As visualized in the dendrogram (see Figure 6), the clustering is derived from co-occurrence patterns in human annotations and captures empirical correlations between emotions. Closely related categories such as joy and excitement, sadness and grief, or anger and annoyance tend to cluster together due to overlapping affective content and shared contextual use. Importantly, these clusters emerged organically in the annotation process without any prior imposition of sentiment groupings, reinforcing the internal coherence and construct validity of the taxonomy. Notably, even ambiguous categories such as realization, curiosity, and surprise form a distinct cluster, which according to Demszky et al. (2020), aligns more closely with positive than negative affect, a pattern suggesting the interpretive subtlety afforded by the taxonomy.

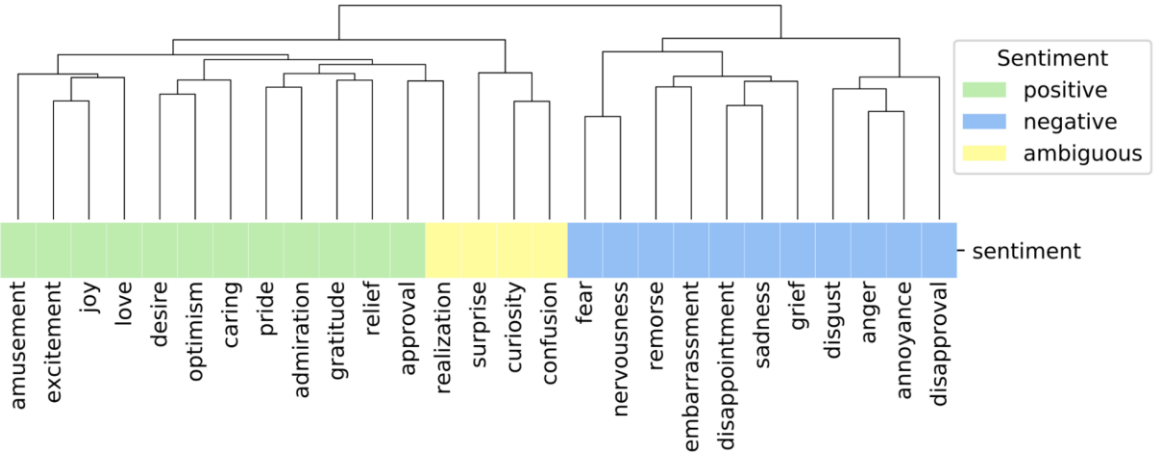

Figure 6: Hierarchical clustering of GoEmotions categories by sentiment orientation and semantic proximity, adapted from Demszky et al. (2020, p. 5)

While GoEmotions supports multi-label annotation to account for the co-occurrence of emotions, this study applies a single-label classification strategy, in which each comment is assigned the most likely emotion category as determined by the classifier. This methodological choice reflects the practical necessity of categorical clarity in statistical modeling, particularly when testing effects on discrete emotional outcomes. Empirically, Demszky et al. (2020) support this approach by reporting that a majority of Reddit comments in the original dataset were annotated with a single dominant emotion, and that inter-annotator agreement was highest for primary labels. This suggests that a single-label strategy is both computationally tractable and aligned with human judgments in emotion labeling for short-form online text.

The GoEmotions taxonomy was operationalized using a BERT-base model fine-tuned on the annotated Reddit dataset, achieving a macro-averaged F1-score of 0.46 across 28 labels (Demszky et al. 2020, p. 7). This performance benchmark establishes BERT as an appropriate baseline model for emotion recognition tasks within informal online discourse. Accordingly, this study adopts a similar BERT-based analysis for the English-language subset of Instagram comments, ensuring methodological consistency while introducing additional robustness through complementary weak labeling using a GPT-based model.

The distribution of annotated examples across emotion categories in the original GoEmotions dataset further complements the taxonomy's conceptual structure. As shown in Figure 7, the frequency of annotations is uneven, with positive emotions such as admiration, approval, and gratitude appearing most frequently. This empirical skew highlights the limitations of traditional emotion classification schemes that focus narrowly on a handful of negative or basic emotions. The prevalence of positive affective language in the Reddit-based corpus underscores the need for a more inclusive taxonomy capable of reflecting the diversity of emotional expression in digital communication (Demszky et al. 2020).

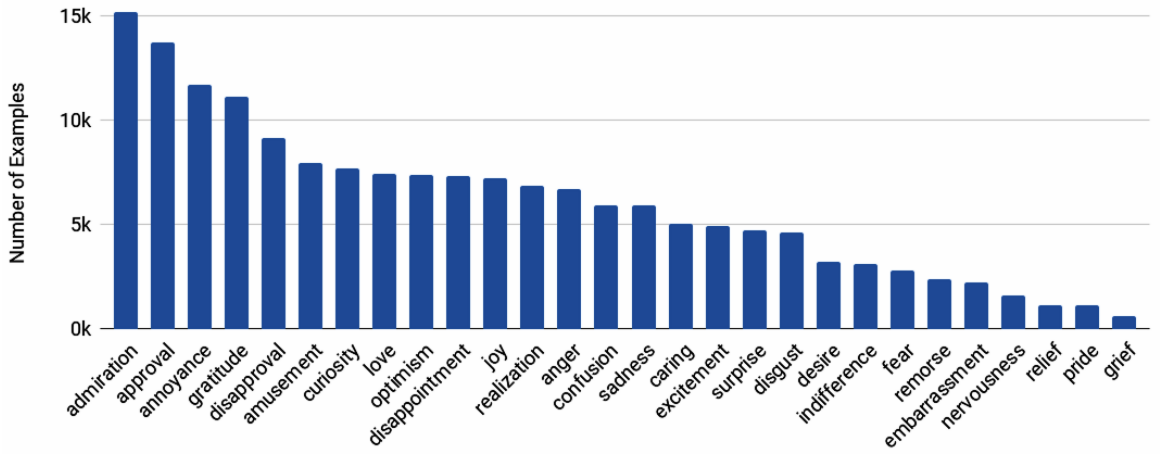

Figure 7: Distribution of annotated emotion categories in the GoEmotions Reddit dataset, adapted from Google Research (2021)

The selection of GoEmotions for this study does not imply a claim to its universal validity, but rather reflects a pragmatic and theoretically grounded methodological decision. In contrast to lexicon-based frameworks or purely dimensional representations, GoEmotions offers a balanced approach in terms of conceptual granularity, empirical rigor, and relevance to informal digital language. Its development from Reddit comments, a corpus that shares key characteristics with Instagram discourse, including brevity, informality, and expressive variability further supports its contextual suitability. As such, the taxonomy provides not only a robust annotation standard but also an analytically coherent foundation for capturing discrete affective responses in computational social science.

4.5.2 ChatGPT for Weak Labeling

The first stage of the classification pipeline uses ChatGPT to assign weak labels across multiple affective and contextual categories, leveraging its scalability for comprehensive coverage of the dataset. Weak labeling refers to the use of indirect, heuristic, or model-based strategies to label data at scale, avoiding the cost and effort of manual annotation while maintaining reasonable labeling fidelity (Ratner et al. 2020). In this context, ChatGPT is tasked with producing structured outputs for each comment, including a primary emotion label from the GoEmotions taxonomy, sentiment polarity, engagement depth, sarcasm classification, and language identification. These outputs serve as foundational variables for downstream statistical modeling and hypothesis testing.

A 1% stratified sample by content origin of comments is extracted from the full dataset to serve as a development set for prompt engineering and model selection. Several LLMs are tested using identical prompts, including Claude 3 Haiku, GPT-3.5-turbo, GPT-4, GPT-4o, GPT-4o-mini and DeepSeek-V3. Models optimized for reasoning are excluded from evaluation due to high costs and limited added value for short-form classification tasks. Few-shot prompting is also tested during development but yields inconsistent results for emotion classification and is hence discarded in favor of a zero-shot design, which proves more stable and better aligned with recent empirical recommendations (Boitel et al. 2024).

The final prompt used in classification is designed to instruct the model to return exactly one label per category, with emotion drawn from the GoEmotions taxonomy (Demszky et al. 2020), along with a model-reported confidence score. A shortened instructional excerpt from the prompt is shown below, with the full version included in Appendix B:

```
You are an expert in analyzing social media comments. Your task is to
analyze a single Instagram comment and classify it into emotional and
engagement-related categories.

Important instructions:

Always return exactly ONE value for "emotion_go". Do NOT return a list or
multiple values.

If multiple emotions seem relevant, choose the single most dominant one.

Definitions:

- emotion_go: [admiration, amusement, anger, annoyance, […]]
```

Figure 8: Snippet of Emotion Classification Prompt

All API requests use a temperature setting of 0.2 to ensure consistent, low-variance predictions across similarly structured inputs. This reduces randomness and encourages deterministic outputs, which is particularly important for classification tasks aiming for replicability (Boitel et al. 2024).

Among the tested models, GPT-4o-mini is selected for full-scale annotation based on its balance of label quality and operational cost. While models like GPT-4o and DeepSeek-V3 exhibit marginal improvements in response quality on the test subset, their token pricing exceeds the proportional benefit for large-scale processing at the time of analysis. The entire dataset is processed via the OpenAI API. Each comment is routed through the final prompt, producing five predicted labels and associated confidence scores. The results are merged with the original metadata, generating a unified dataset for further statistical processing.

Comments labeled as sarcastic, either “Yes” or “Maybe”, are excluded from affective analysis, acknowledging the potential for sarcasm to distort emotion detection (Jahan et al. 2024). Comments identified by the language classifier as “Emoji-Only” or “Emoji-Dominant” are retained in the dataset but explicitly flagged in the language metadata. This design allows the GPT-based emotion distribution to include these comments natively, while still enabling targeted filtering or exclusion in downstream analyses, in recognition of the contextual ambiguity inherent in emoji interpretation (Jahan et al. 2024).

The result is a multilingual dataset, consistently annotated across five analytical dimensions. These labels feed directly into hypothesis testing and serve as the foundation for BERT-based classification and classifier source harmonization.

4.5.3 BERT-Based Emotion Detection

The second stage of the classification pipeline applies a transformer-based model from the BERT family to the subset of Instagram comments previously identified as written in English. This step enhances classification precision by leveraging a model specifically fine-tuned for emotion recognition on the GoEmotions dataset. Whereas the preceding ChatGPT-based approach offers broad coverage across multiple analytical dimensions, the BERT model is employed here for its task-specific optimization and alignment with the GoEmotions taxonomy.

Originally introduced by Devlin et al. (2019), BERT marked a departure from autoregressive architectures like GPT (Radford et al. 2018) by adopting bidirectional pretraining via masked language modeling and next-sentence prediction. Unlike generative models that support general-purpose inference through next-token prediction, BERT is a discriminative model optimized for downstream

classification tasks. Its architecture encodes contextual relationships across entire input sequences, making it especially effective for tasks requiring fine-grained semantic interpretation, such as emotion classification.

While various studies have shown the generalizability of LLMs such as GPT-4 in affective computing (Niu et al. 2024), BERT-based models continue to outperform generative architectures in domain-specific settings where high precision is required (Boitel et al. 2024; Kocoń et al. 2023). This advantage stems from being fine-tuned on labeled corpora, allowing BERT models to internalize task-relevant representational patterns beyond those captured during pretraining. In the context of GoEmotions, BERT serves as the benchmark model against which annotation quality and classifier robustness are evaluated (Demszky et al. 2020). Building on this precedent, the present study adopts a publicly available variant fine-tuned on GoEmotions, roberta-base-go_emotions, distributed under the MIT license and accessible via the Hugging Face Model Hub (Lowe 2024). This RoBERTa-based implementation is structurally similar to BERT but benefits from training on larger corpora and with a modified pretraining objective, enhancing robustness in informal textual domains (Liu et al. 2019).

The model is applied exclusively to the English-language subset of the corpus. This restriction reflects both the monolingual scope of the GoEmotions training data and the observed degradation in classifier performance when applied to multilingual or code-switched input (Niu et al. 2024). While no domain adaptation or additional fine-tuning is performed to account for potential platform-specific divergence between Reddit and Instagram discourse, the risks introduced by this domain shift are explicitly acknowledged as a limitation. Instagram comments may differ from Reddit in syntactic density, expressive norms, and multimodal context, which may reduce the generalizability of the model's learned representations.

Each comment processed by the model is assigned a single dominant emotion label from the 28-category GoEmotions taxonomy, including a neutral class. Although the model internally computes scores for all possible labels, in this study a single-label configuration is used, and only the highest-scoring label is retained. The associated floating-point score, commonly referred to as a probability, is not statistically calibrated but serves as a heuristic confidence measure, consistent with the format used in the GPT-based annotation stage (Lowe 2023). This score is preserved in the metadata schema to support plausibility checks while ensuring methodological consistency across classifiers.

To support initial validation, a stratified subsample of 2% of the English-language subset is manually reviewed to assess plausibility and identify systematic misclassifications. The review sample reflects the underlying distribution of content origin. While this manual assessment does not constitute a full annotation audit, it provides a reference point for identifying edge cases and informs the label harmonization step. The results of this manual validation are reported to assess the plausibility of the classifications and to substantiate the methodological validity of the model's application in this domain.

Through the targeted use of a benchmarked, fine-tuned BERT model, this component of the pipeline complements the breadth-oriented annotation of the GPT stage with a high-precision layer optimized for structured emotion classification. The resulting annotations contribute to a unified dataset architecture in which each instance retains both the source model and confidence score, enabling transparency, robustness checks, and methodological triangulation in downstream analysis.

4.5.4 Label Harmonization and Source Attribution

To ensure consistency across classification outputs, predicted emotion labels are consolidated into a unified metadata structure via a harmonization protocol. Harmonization is limited to standardizing label formatting and recording classifier-specific confidence scores.

Label attribution follows a fixed routing logic: for all English-language comments, labels are sourced from the BERT-based classifier, while the GPT-based annotation is retained for all remaining comments. This decision reflects the complementary roles of the classifiers and is supported by empirical evidence from the validation phase.

To assess the reliability of this configuration, a stratified 2% sample of English-language comments is manually reviewed. Results from this validation, as well as inter-model agreement statistics, are reported to substantiate the adopted pipeline. Classification agreement is quantified using label concordance, defined as the proportion of instances in which both classifiers assign the same label to a given comment (McHugh 2012). This metric provides a transparent basis for evaluating consistency in single-label classification settings where outputs are deterministic.

Additionally, we assess the internal consistency of the emotion classification pipeline by mapping both GPT- and BERT-assigned emotion labels to sentiment polarity (Positive / Ambiguous / Negative), based on the GoEmotions taxonomy introduced by Demszky et al. (2020). Agreement metrics are then computed to evaluate whether classifiers yield convergent polarity-level interpretations.

Each record in the final dataset includes the predicted emotion label, classifier source (GPT or BERT), and the associated probability score. This structure enables transparent analysis and facilitates classifier-aware filtering and validation in downstream modeling.

4.6 Sentiment Classification Approach

In addition to discrete emotion labels, this study incorporates a unified sentiment classification dimension, assigning each comment a polarity label of “positive,” “neutral,” or “negative.” This scalar appraisal of affective tone complements the categorical richness of the GoEmotions taxonomy and supports the testing of Hypothesis 2 (H2), which posits that Content Origin influences the valence of user sentiment. Prior research suggests that users respond more negatively to AI-generated content, even when controlling for aesthetics or thematic content, due to symbolic threat perceptions and authenticity concerns (Bellaiche et al. 2023; Gabbiadini et al. 2024; Park et al. 2024). Sentiment analysis thus serves as an additional lens through which to quantify these evaluative shifts.

Contrasting fine-grained emotion classification, sentiment classification reduces the affective orientation of language to a singular evaluative dimension. This approach is widely adopted in NLP tasks involving user-generated content, particularly in environments marked by brevity, informality, and ambiguity. As D'Andrea et al. (2015) observe, sentiment analysis is particularly well suited for scalable inference in multilingual, low-context settings, where discrete emotional signals may be sparse or ambiguous but general affective tone remains detectable.

Sentiment classification is implemented within the GPT-based zero-shot prompt pipeline. The classification dimension is defined in prompt using an explicit list of the three categories (see Appendix B). This design ensures consistency with the other output dimensions produced by the same model call. The resulting sentiment labels are available across the entire multilingual dataset, including emoji-dominant and informal text formats.

4.7 Engagement Metrics

This study incorporates behavioral engagement metrics to evaluate Hypothesis 3 (H3), which posits that the origin of visual content affects levels of user interaction. Engagement is operationalized using post-level metrics that reflect contribution-type behaviors as outlined in the COBRA framework (Schivinski et al. 2016). Specifically, the number of likes, comments, and reshares attributed to each post serve as

empirical indicators of user response at the contribution tier, corresponding to mid-level engagement that is active but not creative. These metrics are retrieved during the data collection phase via structured API scraping of publicly accessible Instagram post metadata. All post-level fields are directly linked to the post hashtag that indicates content origin.

The primary analytical focus lies on these post-level engagement metrics, as they directly capture interaction with the visual content itself, whose origin is disclosed and classified. Comment-level metrics such as the number of likes or replies received per user comment are also collected. However, as these interactions may reflect engagement with the linguistic content of the comment rather than the original visual stimulus, they are not included in hypothesis testing for H3. Instead, their interpretive function is limited to potential secondary analyses or exploratory use, such as weighting in sentiment aggregation or evaluating the affective salience of certain emotion categories. These exploratory roles could inform future research directions beyond the scope of this thesis.

Post-level engagement metrics exhibit heavy-tailed distributions, a common characteristic of social media data in which a small subset of posts garners a disproportionately large share of user interactions (Bakshy et al. 2012). To provide a robust summary of central tendency under such conditions, this study reports medians rather than means for descriptive tables, as medians are less sensitive to extreme values and offer a more reliable representation of typical engagement (Wilcox 2012). The distributional skew is partly shaped by the data collection design. Posts are retrieved based on their ranking in Instagram's "Top" section. As a result, the dataset reflects posts that have already achieved elevated visibility and interaction. This introduces a selection bias toward highly engaged content. The study does not seek to estimate overall engagement across all content, but to evaluate engagement patterns among prominent posts within each content origin category. Accordingly, the engagement metrics used are representative of high-visibility examples of AI-generated and human-created content, rather than of the entire distribution of published material.

For the purpose of hypothesis evaluation, engagement data is aggregated by content origin and grouped by hashtag labels. While raw values provide a comparative lens, the differing volume of posts across origin groups necessitates normalization or adjustment through statistical modeling to ensure interpretive comparability.

4.8 Engagement Depth Classification

This study incorporates engagement depth as an additional indicator of user response, enabling evaluation of Hypothesis 4 (H4), which posits that the origin of visual content influences the depth of user engagement. Defined as the level of cognitive and emotional elaboration expressed in user comments, engagement depth extends the analytical scope beyond surface-level frequency metrics. Prior studies suggest that AI-generated content is perceived as less emotionally authentic and interpretively rich, potentially resulting in shallower commentary (Bauer et al. 2024). The classification of engagement depth supports this hypothesis by capturing variations in user elaboration associated with content origin.

Engagement depth is operationalized as an ordinal variable, assigned at the comment level through a GPT-based classifier. Each comment is categorized into one of four engagement depth levels: Superficial, Moderate, Deep, and Very Deep. These categories represent increasing levels of interpretive effort and emotional involvement, from minimal engagement (e.g., emoji-only responses or generic praise) to complex, reflective expressions.

```
Definitions:
   -  […]
   -  engagement_depth:
      Superficial: short, emoji-only, or generic
      Moderate: brief, specific compliment
      Deep: thoughtful, critique, or question
      Very Deep: personal story, emotional reflection
```

Figure 9: Snippet of Engagement Depth Classification Prompt

Classification is conducted within the GPT-based pipeline stage. The engagement depth dimension is included as part of the same prompt that returns emotion, sentiment, sarcasm, and language labels, ensuring consistency in the annotation process. The final prompt is implemented using a few-shot strategy, which includes explicit category definitions (see Figure 9) and representative examples. This design choice follows systematic testing on the 1% stratified sample of comments comparing zero-shot and few-shot variants of the prompt. Few-shot prompting is selected based on its superior ability to differentiate between semantically similar categories, particularly between Deep and Very Deep responses, while maintaining alignment with defined criteria. To illustrate the model's interpretive guidance, Table 4 presents the representative example per category, drawn directly from the finalized few-shot prompt.

Category	***Example Text***
Superficial	"🔥🔥🔥🔥🔥"
Moderate	"It looks great! I love the cools in the lightning"
Deep	"What white pen do you use. I've been looking for a good one for a long time"
Very Deep	"You are a huge inspiration. I am from right outside […] but now live in […]. I'm tight with the guys at […] and they said you came in a while back. I was in high school when I purchased your paperback (with your cover). Yourself and many others have influenced me since getting back into painting about six years ago. […] So funny how these things work out. Take care!"

Table 4: Few-Shot Comment per Engagement Depth Category

The resulting engagement depth labels are stored alongside each comment and used as an ordinal outcome variable in the statistical testing of Hypothesis 4. This variable complements the frequency-based behavioral metrics by introducing an interpretive dimension of user response. While likes and comment counts measure the extent of interaction, engagement depth captures its expressive character, providing additional insight into how users respond to AI-generated versus human-created content.

4.9 User Segmentation by Exposure Pattern

This section outlines how users are segmented to operationalize the moderator variable User Exposure Pattern, as introduced in the conceptual model (see Chapter 3.3). This segmentation forms the analytical foundation for testing Hypothesis 5 (H5), which posits that users exposed to both AI-generated and human-created content exhibit more differentiated emotional responses than those exposed to only one content type.

While the dataset is structured at the comment level, segmentation occurs at the user level, based on the anonymized userID field. For each user, interactions with posts tagged under #aiart and #traditionalart are analyzed to infer prior exposure. Users who comment at least once under both hashtags are classified as having mixed exposure. All others are assigned to one of two mutually exclusive categories: AI-only (engaged exclusively with #aiart content) or traditional-only (engaged exclusively with #traditionalart content). This classification produces a categorical variable called User Exposure Pattern, which is stored alongside each unique userID in the analytical dataset.

A single comment under each hashtag type is sufficient for a user to qualify as mixed. This inclusive approach ensures broad coverage of exposure histories while maintaining interpretive clarity. As the segmentation relies exclusively on public, anonymized behavioral data, it adheres to privacy-by-design and complies with ethical research standards.

The theoretical rationale for this segmentation draws on constructivist emotion theory and appraisal theory, which emphasize that emotional evaluations depend on contextual interpretation and prior knowledge (Barrett 2006; Smith and Lazarus 1993). By identifying users with broader exposure histories, the study enables comparative analysis of how familiarity with both content types may moderate affective responses. While the segmentation itself does not imply a directional effect, it serves as a key analytical dimension for modeling potential interaction effects on emotional outcomes.

4.10 Variable Operationalization

This section summarizes the operational definitions and computational extraction procedures used to transform raw social media data into analyzable variables. Each variable corresponds to a conceptual construct outlined in the research model. Operationalization includes both variable definition and the specification of measurement methods. Classification outputs are produced by automated NLP pipelines, primarily relying on GPT-based weak labeling and, in the case of English-language emotion labels, BERT-based refinement.

The independent variable Content Origin is derived from post-level hashtag annotations. Posts tagged with #aiart are classified as AI-generated, while those tagged with #traditionalart are coded as human-created. Only posts with clearly interpretable and explicit origin cues are included in the dataset. The resulting binary variable is assigned at the post level and linked to each associated comment to enable consistent comparison across user responses.

Transparency is embedded as a control condition within the dataset. All AI-generated content analyzed in this study is transparently labeled as such by the content creator through the inclusion of the #aiart hashtag. Because this disclosure is a selection criterion and does not vary within the dataset, Transparency is not modeled as a separate predictor. However, its conceptual relevance remains critical, as it informs the interpretive framing of Content Origin and interpretation through Supporting Hypothesis 1 (HS1).

The dependent variable Emotion, used to test differences in affective response, is defined according to the GoEmotions taxonomy (Demszky et al. 2020), which includes 27 discrete emotion categories and a neutral class. Emotion labels are assigned through a dual-stage classification pipeline, consisting of a zero-shot GPT-based annotation applied to all comments and a BERT classifier used on the English-language subset. Each comment receives one dominant emotion label. Metadata includes classifier source and confidence scores, allowing for robustness checks and model transparency.

Sentiment is operationalized as a categorical evaluation of general affective tone and is classified within the same GPT-based prompt used for emotion detection. Each comment is labeled as Positive, Neutral, or Negative. Sentiment classification is designed for linguistic generalizability and complements the emotion taxonomy by capturing valence in a coarse but interpretable manner. To assess internal consistency, sentiment polarity is compared post hoc to the valence groupings of emotion labels.

Three Engagement metrics (number of likes, comments, and reshares) are recorded at the post level. These metrics are collected via structured scraping of publicly accessible metadata and directly reflect user interaction with visual content. Because the dataset includes only highly ranked “Top” posts, these figures represent engagement with prominently visible content rather than a representative sample of overall platform activity. Each metric is linked to post-level content origin and used to evaluate behavioral response patterns.

Engagement Depth captures the qualitative richness of user responses and is defined as a categorical variable reflecting the degree of elaboration in comment text. Each comment is categorized into one of four levels: Superficial, Moderate, Deep, or Very Deep. The classification is generated via a GPT-based few-shot prompt, which includes explicit definitions and illustrative examples. This variable introduces an interpretive dimension to user interaction and complements the frequency-based engagement metrics by capturing expressive intensity.

The variable User Exposure Pattern functions as a moderator in the analysis. It is derived by analyzing user-level behavior based on anonymized userIDs. Users are categorized into three groups: AI-only (engaged exclusively with AI-generated content), Traditional-only (engaged exclusively with human-created content), and Mixed (engaged with both types). This categorical variable reflects differential exposure histories and enables the analysis of interaction effects with Content Origin on emotional outcomes.

Two additional variables are produced as part of the GPT-based annotation pipeline but are not directly entered into the analytical models. Language is inferred for each comment and used to determine eligibility for secondary classification by BERT. English-language comments are routed through both GPT and BERT pipelines, while all others are retained with GPT-only annotations. Sarcasm is also detected via GPT. Rather than being used as a modeling variable, sarcasm detection is employed as a conservative filtering mechanism: comments flagged as sarcastic, either definitively or tentatively, are excluded from emotion and sentiment analyses to preserve classification validity and interpretive reliability.

Table 5 provides an overview of all variables used in statistical modeling, including their data type, computational source, and preprocessing steps. It includes variables that serve as predictors, outcomes, or moderating factors in hypothesis testing.

Variable	*Type*	*Measurement / Labeling Source*	*Preprocessing / Notes*
Content Origin	Binary	Hashtag-based (#aiart, #traditionalart)	Only posts with clearly labeled origin are included.
Transparency	Binary (control)	Implicit via hashtag inclusion	All AI content is included only if transparently disclosed by the creator; variable does not vary and is not modeled.
Emotion	Categorical (28 classes)	GPT (multilingual), BERT (English)	Harmonized to GoEmotions taxonomy. BERT overrides GPT for English.
Sentiment	Categorical (positive, neutral, negative)	GPT	Derived from GPT zero-shot label; cross-validated with emotion polarity.
Engagement Metrics	Continuous	Post metadata (API)	Likes, comments, reshares.
Engagement Depth	Categorical (4 classes)	GPT (few-shot prompt)	Labels: Superficial, Moderate, Deep, Very Deep.
User Exposure Pattern	Categorical (3 classes)	Derived from userID behavior	AI-only, Traditional-only, or Mixed based on comment history.
Language	Categorical	GPT	Used for routing to BERT; flagged for emoji-dominance.
Sarcasm	Categorical (Yes/Maybe/No)	GPT	Labeled for exclusion; comments flagged as sarcastic are removed from affective analyses.

Table 5: Operationalization of Variables

4.11 Statistical Analysis

This chapter outlines the statistical procedures used to evaluate the study's five primary hypotheses (H1–H5), grounded in the analytical framework established. All analyses are pre-specified and selected to ensure robust estimation, transparent assumptions, and appropriate treatment of the dataset's scale and structure. Adjustments are permitted only in cases of significant model assumption violations and must remain within the same statistical family, in line with methodological guidance by Recker (2021).

In accordance with this framework, statistical tests are chosen based on the nature of the variables involved and implemented using Python with specialized libraries. Where appropriate, model diagnostics, effect size measures, and post hoc comparisons are employed to ensure analytical rigor and interpretive clarity. The analytical approach follows principles outlined by Agresti (2018), Cameron and Trivedi (2013), Hilbe (2011), and Wilcox (2012).

Because our focal construct is the overall distribution of audience emotions by content origin, all χ^2 tests are computed on the full comment pool, treating each comment as an independent sample conditional on origin. The large number of distinct posts and users makes any residual intra-class correlation more likely to attenuate effect sizes than to inflate Type I error. Hence, the results should be interpreted as average marginal tendencies rather than post- or user-specific effects.

4.11.1 Emotion Distribution (H1)

This test evaluates whether the distribution of user-reported emotions differs between comments on AI-generated versus human-created artwork. The dependent variable is the dominant emotion assigned to each comment, categorized according to the 28 GoEmotions labels (Demszky et al. 2020). To estimate the effect of content origin on emotional responses, a baseline-category multinomial logit (MNL) model is employed. The reference category is neutral, and for each other emotion category $c \in \{1, \ldots, 27\}$, the model estimates the log-odds of observing category c relative to neutral as a linear function of origin:

$$log(^{P(Y_i = c)}/_{P(Y_i = neutral)}) = \beta_{0c} + \beta_{1c} \times Origin_i$$

where:

Yi is the dominant emotion label for comment i,

$Origin_i \in \{0,1\}$ denotes the content origin (Human = 0, AI = 1),

And β_{1c} captures the log-odds shift for emotion category c relative to neutral.

Hypothesis testing is conducted using Wald χ^2 statistics to jointly assess the null hypothesis that content origin has no effect on emotional expression:

$$H_0: \beta_{1c} = 0 \; for \; all \; c \neq neutral$$

For each emotion category, the effect size is interpreted via the odds ratio (OR), calculated as:

$$OR_c = exp(\beta_{1c})$$

This quantifies the multiplicative change in the odds of observing emotion c (versus neutral) when content origin shifts from Human to AI. To support interpretation, average predicted probabilities (APPs) for each emotion are computed by origin group and visualized with 95% bootstrapped confidence intervals (CI). Emotion categories with sparse occurrence counts are monitored, and coefficients for categories with insufficient support (e.g., $n < 10$) are flagged. If necessary, such categories are collapsed or removed to maintain model stability (Agresti 2018).

4.11.2 Sentiment Polarity (H2)

To evaluate whether sentiment polarity differs by content origin, a Pearson Chi-square test of independence is performed on a 2 × 3 contingency table, crossing Content Origin (AI vs. Human) with Sentiment (Negative / Neutral / Positive). The test statistic is computed as:

$$\chi^2 = \sum_{r=1}^{2} \sum_{c=1}^{3} {(O_{rc} - E_{rc})^2}/{E_{rc}}, \qquad with \qquad E_{rc} = {(row_r)(col_c)}/{N}$$

where O_{rc} and E_{rc} denote the observed and expected frequencies for cell (r, c), respectively, and N is the total number of observations. The resulting statistic follows a χ^2 distribution with 2 degrees of freedom under the null hypothesis of independence, provided that at least 80% of expected counts exceed 5 (Agresti 2018). The null hypothesis tested is:

$$H_0: Sentiment \; is \; independent \; of \; Origin$$

The effect size is reported using Cramér's V, a normalized measure defined as:

$$V = \sqrt{{\chi^2}/{N(k-1)}}, \qquad where \qquad k = min(2,3) = 2$$

A 95% confidence interval around Cramér's V is also reported. If the omnibus χ^2 test is significant, standardized residuals are inspected to identify which sentiment categories deviate from expectation (Wickens 2014). To isolate specific contrasts, Holm-corrected pairwise 2 × 2 χ^2 tests are applied (Wilcox 2012), controlling for multiple comparisons.

To verify the construct validity of the sentiment classification, GPT-derived Sentiment labels are compared against valence codes mapped from the 28 GoEmotions categories (Positive / Neutral / Negative). A second Pearson χ^2 test assesses distributional independence, and Cohen's κ quantifies classification agreement. Values of $\kappa \geq .80$ are interpreted as strong concordance (McHugh 2012), and

residuals are reviewed to detect systematic mismatches. This cross-check strengthens confidence in the sentiment classification as a valid proxy for emotional valence.

Together, these analyses evaluate whether content origin systematically shifts the affective polarity of audience reactions and confirm that the sentiment variable is internally consistent with the emotion-based framework introduced in H1. As established, results reflect average marginal tendencies across the deduplicated dataset.

4.11.3 Engagement Volume (H3)

To test whether the volume of behavioral engagement differs by content origin, separate Negative Binomial Regression (NBR) models are estimated for three dependent variables: Likes, Comments, and Reshares. Each is measured as a count variable, with Content Origin (AI vs. Human) serving as a binary predictor. Unlike H1–H2, the unit of analysis here is the post, because engagement counts accrue to the post rather than the individual comment. Given the observed right-skew and overdispersion typical of social media engagement data, NBR is preferred over Poisson regression due to its inclusion of a dispersion parameter and greater robustness to variance inflation (Cameron and Trivedi 2013; Hilbe 2011).

Formally, let Y_i be the observed count for post i with mean μ_i. A Poisson model assumes $Var(Y_i) = \mu_i$, whereas the negative binomial generalizes this with an overdispersion parameter $\alpha > 0$, such that:

$$Var(Y_i) = \mu_i + \alpha\mu_i^2$$

The mean μ_i is modeled via a log link function:

$$log(\mu_i) = \beta_0 + \beta_1 \times Origin_i, \qquad with \qquad Origin_i \in \{0,1\}\,(Human = 0,\ AI = 1)$$

Maximum-likelihood estimation yields $\widehat{\beta_1}$, whose exponentiation gives the incidence-rate ratio (IRR):

$$IRR = exp(\widehat{\beta_1})$$

This quantity reflects the multiplicative change in expected engagement volume when the content origin shifts from human to AI.

Each NBR model reports model-based predicted means with 95% confidence intervals to aid interpretation. Overdispersion is quantified using the model's theta (θ) parameter, and Wald z-tests are used to evaluate the statistical significance of the origin coefficient. Model assumptions are validated via residual diagnostics, which also help identify potential outliers, especially viral posts with high leverage.

If structural zeros account for more than 20% of observations for any outcome, most plausibly in the case of reshares, a zero-inflated negative binomial model is estimated as a robustness check (Agresti 2018).

This modeling approach provides a distributionally appropriate test of whether content origin predicts variation in engagement volumes across distinct user behaviors. It complements the categorical analyses in H1 and H2 by targeting continuous response patterns in user interaction data while maintaining methodological alignment with prior sections.

4.11.4 Engagement Depth (H4)

To assess whether the distribution of engagement depth differs between AI- and human-generated art, a Pearson χ^2 test of independence is conducted on a 2 × 4 contingency table, crossing Content Origin (AI vs. Human) with Engagement Depth, categorized into four qualitatively distinct levels (Superficial/

Moderate/ Deep/ Very Deep). These categories represent distinct degrees of user interaction and are treated as nominal.

The statistical procedure, including the formulation of the χ^2 statistic, distributional assumptions, and effect size reporting via Cramér's V, follows the definition introduced in 4.11.2 (Agresti 2018). Specifically, the test evaluates whether engagement depth is independent of content origin and reports a $\chi^2(3)$ statistic. Assumption checks ensure that at least 80% of expected cell counts exceed 5, in line with standard recommendations (Field et al. 2012; McHugh 2012). Where this criterion is met, significance is interpreted using the omnibus test and standardized Pearson residuals identify individual depth levels that contribute most strongly to the observed association (Wickens 2014).

To complement the χ^2 test and provide model-based effect sizes, a baseline-category multinomial logit model is estimated with Engagement Depth as the outcome and Content Origin as the sole predictor. This procedure, previously defined in Chapter 4.11.1 (Agresti 2018), treats the Superficial category as the reference. The model estimates the log-odds of deeper engagement levels as a function of origin, and odds ratios quantify the multiplicative change in likelihood when comparing AI- to human-generated content. In combination, the chi-square and multinomial analyses provide a methodologically coherent test of whether content origin shapes how deeply users engage across distinct categories of interaction.

4.11.5 User Exposure Pattern (H5)

This hypothesis tests whether the emotional effect of content origin is moderated by users' exposure pattern. Specifically, it evaluates whether users who engage with both AI and traditional artworks ("Mixed") respond differently than those exposed to only one content type ("AI-only" or "Traditional-only").

A baseline-category multinomial logit model is fit with L_2 regularization to account for sparse cells, particularly in the Mixed group. Let $\mathrm{Origin_i} \in \{0,1\}$ (Human = 0, AI = 1), and define two exposure indicators: $\mathrm{D_{i1}} = I[Expo_i = \mathrm{AI\text{-}only}]$, and $\mathrm{D_{i2}} = I[Expo_i = \mathrm{Mixed}]$. For each non-neutral emotion c, the log-odds are modeled as:

$$log(^{P(Y_i = c)}/_{P(Y_i = neutral)}) = \beta_{0c} + \beta_{1c} \times Origin_i + \sum_{\mathrm{k=1}}^{2} [\beta_{2c,k} \times \mathrm{D_{ik}} + \beta_{3c,k} \times Origin_i \times \mathrm{D_{ik}}]$$

To test moderation, we evaluate the joint null hypothesis $H_0: \beta_{3c,k} = 0 \; for \; all \; c \neq neutral$ via a bootstrap-based Wald χ^2 test. We draw 1,000 stratified resamples, re-fit the ridge-MNL, and compute empirical p-values based on the resulting χ^2 distribution.

For each exposure group $\mathrm{g} \in \{AI - only, Mixed\}$, interaction effect sizes are reported as odds ratios $OR_{g,c} = exp(\beta_{3c,g})$ with 95% bootstrap percentile confidence intervals. Holm-adjusted p-values account for multiple comparisons.

To illustrate practical differences, APPs are calculated across all Origin × Exposure combinations. Emotion categories with zero support in any exposure group are excluded, others are retained regardless of frequency due to model regularization.

This specification extends the categorical modeling logic of H1 while ensuring valid inference under sparsity via regularization and bootstrapping (Agresti 2018; Hilbe 2011; Recker 2021).

4.11.6 Summary of Statistical Procedures

The table below summarizes the statistical tests and variables implemented to test hypotheses H1-H5, providing a consolidated overview of the analytical design. All results are interpreted as average marginal tendencies drawn from the deduplicated dataset.

Hypothesis	*Variables (IV → DV)*	*Primary Test / Model*	*Effect Size metric (primary)*
H1	Content Origin → Emotion	Baseline-category MNL	OR (per emotion)
H2	Content Origin → Sentiment	Pearson χ^2 (2 × 3)	Cramér's V
Validation	Sentiment ↔ Emotion Valence	Pearson χ^2	Cohen's κ
H3	Content Origin → Likes / Comments / Reshares	NBR (three models)	IRR
H4	Content Origin → Engagement Depth	Pearson χ^2 (2 × 4)	Cramér's V (supplementary ORs from MNL)
H5	Content Origin × Exposure Pattern → Emotion	Ridge-penalized MNL	Interaction ORs

Table 6: Statistical Procedures

4.11.7 Overview of Analytical Environment

All analyses in this study are implemented in a Python 3.12 environment configured for reproducibility, scalability, and integration with both statistical and NLP pipelines. Package versions are pinned throughout to ensure stable outputs across iterations of modeling, annotation, and evaluation.

The software architecture supports all core procedures outlined in preceding sections, including dual-stage emotion classification, regularized MNL estimation, and bootstrapped statistical inference. Annotation results are stored in intermediate parquet files and incrementally merged into a harmonized analysis dataset. All modeling inputs, outputs, and diagnostics are stored to ensure transparency and auditability.

Table 7 summarizes the libraries and frameworks used, organized by analytical function and aligned with the methodological steps they support. The exact library versions used are documented in a reproducibility-focused requirements.txt file, included in Appendix C.

Component	*Library*	*Application*
Data Handling & Numerics	pandas	Structuring and transforming tabular data
	numpy	Numerical operations, matrix transformations
	scipy	Bootstrap routines and distributional inference
NLP Classification	transformers	Loading and applying BERT-based emotion classifier
	huggingface-hub	Retrieval and versioning of fine-tuned BERT model
	torch	Backend execution for transformer inference
	openai	GPT-4o-mini API calls for zero-shot weak labeling
Statistical Modeling	statsmodels	Estimating MNL, negative binomial, and zero-inflated models
	scikit-learn	Ridge-penalized logit models, stratified bootstrapping

Table 7: Core Libraries and Packages

This configuration ensures coherence between statistical estimation and annotation logic, with model outputs feeding directly into hypothesis testing. All code is executed with deterministic seeds, and runtime metadata is archived to facilitate replication and future robustness checks.

4.12 Methodological Limitations

This chapter outlines the primary methodological limitations of the study, reflecting constraints in data collection, annotation quality, model assumptions, and interpretive scope. While the analyses offer meaningful insight into how audiences engage with AI-generated versus traditional artwork, they do so under conditions that reflect the present state of generative AI, platform-specific structures of Instagram, and the available computational tools. The following limitations define the boundaries within which the findings should be interpreted.

4.12.1 Data Collection and Sampling

The dataset is sourced exclusively from the Top tab of selected hashtags on Instagram, capturing highly visible and algorithmically boosted content. As a result, the analysis generalizes to high-engagement art posts but does not represent the platform's long-tail of lower-visibility or niche content.

Data collection spans a fixed one-month window (8 March - 8 April 2025), introducing potential short-term fluctuations or seasonal effects. In the evolving domain of generative AI, where adoption and perception shift rapidly, the findings reflect a specific moment in time rather than a stable or enduring pattern of interpretation.

Only the first page of comments per post is available for analysis, which limits access to deeper or more contentious sub-thread discussions. This constraint yields a broad overview but not a complete picture of user emotion in comment sections.

The classification of content origin relies on the presence of hashtags (e.g., #aiart). Posts without such tags are excluded and mislabeling may occur in both directions. The analysis presumes that users notice and interpret these tags as origin signals, but no observational or survey data confirm this assumption. In practice, hashtags are positioned below content descriptions and often remain collapsed by default, raising uncertainty about user exposure to the label.

The operationalization of User Exposure Pattern (AI-only, Traditional-only, Mixed) is restricted to interactions within the collected dataset. User activity outside the one-month window or on other platforms remains unobservable.

4.12.2 Measurement and Annotation

Emotion and sentiment labels are generated through a hybrid GPT-BERT classification pipeline. The BERT model, trained on Reddit and evaluated on GoEmotions, achieves a macro-F1 score of approximately 0.46, and may underperform on Instagram due to differences in language style, tone, and platform norms. The GoEmotions corpus itself introduces known structural biases that constrain generalizability. Its Reddit-based sample under-represents emotional phrasing common among female, multilingual, and non-Western users, while profanity filtering reduces the model's sensitivity to strong negative affect. Annotation by native English speakers in India may further limit cross-cultural emotion recognition, especially for dialects and pragmatic cues common on Instagram (Demszky et al. 2020).

The pipeline introduces label noise and systematic category misclassification. These inaccuracies may bias category-specific effect estimates. The pipeline also enforces a single dominant emotion per comment, which discards co-occurring emotions and potentially oversimplifies complex affective expression.

Non-English comments are retained where GPT assigns a label, but cross-linguistic differences may result in semantic misinterpretation. Emotion and sentiment classifications for these comments originate from the same pipeline, which may inflate κ statistics in validation tests due to shared error sources.

Sarcasm detection is implemented through a simple GPT-based flag, which may miss subtle or culturally embedded irony, leaving residual contamination in emotion estimates.

4.12.3 Model Assumptions and Statistical Inference

All models assume independence of observations, yet comments are nested within posts and users. This nesting structure is unmodeled and intra-class correlations are not reported. Standard errors may therefore understate true variance.

To address sparse category counts, emotion labels with fewer than 10 observations are collapsed or removed. This rule is defined ex ante but executed data-dependently, which may affect type I error rates and model sensitivity.

The study includes a high number of simultaneous statistical tests, 27 for H1 and 54 for H5. Although Holm correction is applied, the risk of family-wise error inflation remains nontrivial.

In H3, the use of Negative Binomial Regression assumes a constant dispersion parameter (α). If over-dispersion varies substantially across posts, content types, or hashtag clusters, the assumption may be violated. For outcomes like Reshares, which may display structural zeros, zero-inflated models are employed where necessary, but residual model misfit may persist.

Regarding H4, engagement depth is treated as nominal, avoiding the imposition of an ordinal structure. While this conservatively respects the informality of social-media engagement, it also sacrifices statistical power and precludes the use of ordered-logit models.

The three-way MNL model used in H5 introduces additional complexity and parameter expansion. Infrequent combinations of exposure group and emotion category result in sparse cells, which may destabilize estimation and limit interpretability.

4.12.4 External Validity and Interpretation

All findings are derived from data collected on Instagram, a platform that prioritizes visual engagement, aesthetic signaling, and algorithmic curation. The results may not generalize to text-oriented or less visually driven platforms.

The study adopts a cross-sectional design, offering a snapshot of behavior rather than insight into longitudinal dynamics, habit formation, or causal change in user sentiment.

Behavioral metrics serve as observable proxies for user engagement, but they do not directly measure internal attitudes. Private interactions, shares or prolonged viewing are not captured. While emotion classification enhances interpretability, it cannot fully resolve this inferential gap.

Finally, the study lacks demographic metadata. User characteristics such as age, geography or artistic expertise may moderate the effect of content origin, but they remain unobserved due to platform limitations and data protection considerations.

4.12.5 Technical Constraints

The dataset is collected using a third-party scraping interface, which operates outside the official Instagram API. Although effective at the time of collection, this method may become unavailable due to platform updates or enforcement of Terms of Service, posing risks to replication and continuity.

Manual validation is conducted on a stratified audit sample, which supports initial pipeline assessment but lacks sufficient scale to detect systematic labeling errors across emotion categories or languages.

4.13 Ethical Considerations

This study involves large-scale computational analysis of user-generated content from Instagram. While all comments are publicly accessible and collected without authentication, ethical considerations arise concerning data protection, user expectations, affective inference, and algorithmic responsibility. The research protocol adheres to the standards of computational social science, complies with the General Data Protection Regulation.

Data collection proceeds under the lawful basis of legitimate interests (European Union 2016, Article 6(1)(f)). User identifiers are irreversibly hashed using a salted SHA-256 function, and no re-identification key is retained. Raw comments are stored in pseudonymised form, separated from metadata on encrypted, access-restricted infrastructure. Although fully anonymised data is not achieved under GDPR definitions, identifiability is minimized through technical safeguards and strict access control. Raw comment text is retained for a maximum of twelve months, solely for reproducibility and audit, after which it will be securely deleted. While the automated collection of Instagram comments formally violates the platform's Terms of Use, current scholarship recognises such scraping as ethically permissible when limited to public content, conducted without authentication, and rate-limited for academic research (Jünger 2023). In line with the principle of contextual integrity (Nissenbaum 2010), the reuse of publicly visible comments is considered appropriate. The data is shared in a public forum, analysed exclusively in aggregate, and never quoted verbatim. The dataset is limited to fields required for hypothesis testing. No images, location data, biographies, or follow relationships are retained. All data handling complies with GDPR on data minimisation (European Union 2016, Article 5(1)(c)), ensuring that only the minimum necessary information is collected and processed for the research objectives.

Emotion and sentiment are inferred using a hybrid GPT-BERT pipeline. As discussed in Chapter 4.12.2, these models are trained on data with known demographic and linguistic biases and may fail to capture dialectal variation or culturally embedded affect. Profanity filtering in the GoEmotions training corpus further limits sensitivity to strong negative emotions. Model outputs are therefore used for aggregate pattern analysis only, and no individual-level inference is attempted.

Given that affective inference may involve sensitive psychological attributes such as emotional tone, the study refrains from releasing any raw or pseudonymised comments. Instead, outputs are restricted to statistical summaries and synthetic or masked exemplars.

Taken together, these measures aim to ensure that the study remains transparent, ethically sound, and privacy-respecting, while enabling reproducible and socially beneficial research.

4.14 Methodological Reflections and Trade-offs

This chapter reflects on the methodological design decisions, considering the logic behind key trade-offs and what they imply for the evidential scope of the study. Rather than reiterating individual limitations, we clarify how competing priorities, scalability, interpretability, cultural fit and legal compliance, are balanced in constructing the pipeline.

The pipeline's primary strength lies in its scalability: affective and behavioral labels are extracted for tens of thousands of comments at low cost. This throughput enables statistical modeling with rich covariates but comes at the price of interpretability. Each label is model-dependent, and LLM decisions are not fully auditable. We mitigate this opacity by storing model confidence scores, manually auditing a subsample, and triangulating results between GPT and BERT, but a formal error-propagation model remains a priority for future research.

Breadth is achieved through multilingual, zero-shot annotation, which captures non-English and emoji-rich expressions across Instagram's global user base. However, the classification taxonomy is English-centric and BERT only processes English text. This creates a mismatch between global language diversity and the tools that are used to model affective response. A language-adaptive pipeline, trained on region-specific corpora, could improve cultural fit in future replications.

The legal and ethical design prioritises GDPR compliance over analytical completeness. Hashing, data minimisation, and a retention limit reduce identifiability risk but also eliminate contextual metadata such as profile language or network size that might enrich interpretation.

The study is hypothesis-driven in structure but exploratory in execution. Operational constructs such as user exposure and engagement depth are novel and platform specific. This hybridity enables responsiveness to real-world data but limits construct validity, especially for engagement depth, which lacks external validation. A priority for future work is to validate the ordinal assumptions behind the four-level depth scheme, for instance by comparing it to user-rated scales.

Methodologically, the architecture is robust. Its modular structure allows for component-wise updates, from scraper to classifier to statistical model, without redesigning the entire workflow. The ethics-by-design framework based on GDPR principles, pseudonymisation, and transparent variable selection scales effectively without introducing administrative or computational overhead.

In sum, the approach reflects a calculated balance: scalable affect mining and ecological realism constrained by partial interpretability, cultural specificity, and causal certainty. This trade-off is justified by the scope and intent of the study but also defines the boundaries of inference. Future refinements should focus on cultural adaptation, human-in-the-loop validation, and statistical modeling that captures the nesting and temporality of social media discourse. This chapter closes the methodological arc of the thesis, situating the results that follow within a clearly bounded evidential frame.

5 Results

This chapter presents the empirical results of the study, organized by the five research hypotheses (H1–H5) introduced in Chapter 3 and operationalized in Chapter 4. Findings are structured by hypothesis and supported by descriptive statistics, model estimates, and visual summaries where relevant. The results are based on a harmonized dataset of 64,806 user comments, annotated for emotion, sentiment, and engagement attributes through a dual-stage NLP pipeline. Statistical procedures match the data structure for each hypothesis and include chi-square tests, count models, and multinomial logistic regression.

5.1 Descriptive Summary of the Dataset

This section presents a descriptive summary of the dataset used for primary and comparative analyses. It outlines the final sample size after filtering as well as the distribution of comments by content origin and language.

The dataset used for hypothesis testing consists of user comments posted under the hashtags #aiart and #traditionalart. Following removal of sarcastic or pinned comments, invalid GPT-classified emotion categories, and duplicate entries, the final multilingual sample includes $N = 64{,}806$ comments.

Raw Data	***65,843***
Sarcasm & isPinned Filter	970
Invalid emotion classification	55
Deduplication	12
Final Dataset	64,806

Table 8: Dataset Overview

Of the multilingual sample, $N_E = 14{,}293$ comments are classified as English-language and serve as the basis for BERT-based classification.

	aiart	***traditionalart***
Multilingual Dataset	43,541 (67.2%)	21,265 (32.8%)
English subset	7,142 (50.0%)	7,151 (50.0%)

Table 9: Content Origin Distribution

To support classifier evaluation, a stratified 2% audit sample was drawn from each origin group, yielding 143 comments from #aiart and 144 from #traditionalart.

The final multilingual dataset includes 64 identified languages, including emoji-only and emoji-dominant classification. The most common languages within each origin category are summarized below, with less common languages aggregated into “Other”.

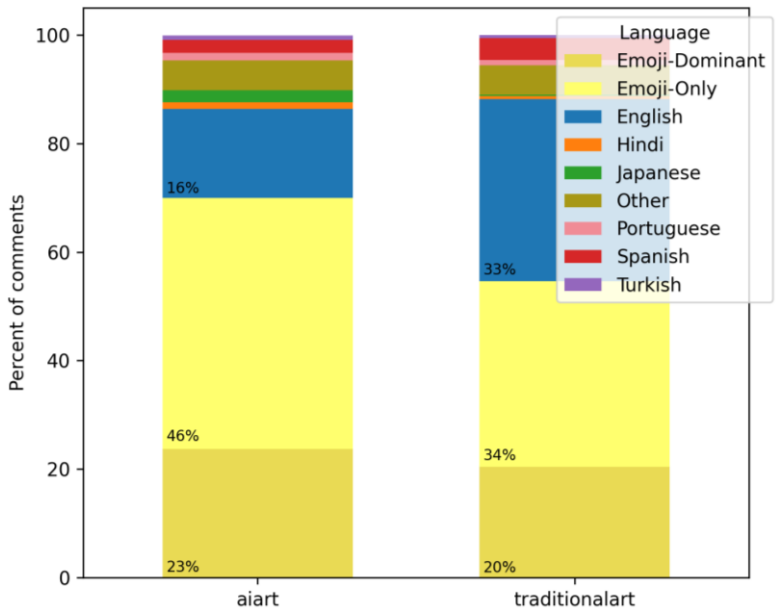

Figure 10: Language distribution by content origin

Emoji-based and multilingual comments are included in the main analyses and annotated using the GPT-based classifier. Only English-language comments are passed through the secondary BERT classifier.

5.2 Classifier Concordance and Validation

This chapter summarizes validation procedures conducted to assess the agreement between classifiers and the internal consistency of their output. While not part of hypothesis testing, these steps enhance the transparency and robustness of the emotion classification pipeline. Validation is conducted through three complementary procedures: a manual audit of annotation plausibility, automated label comparison between GPT and BERT classifiers, and polarity-level concordance checks based on mapped sentiment categories.

5.2.1 Plausibility Audit of Classifier Outputs

To assess the face validity of the emotion labels assigned by each classifier, a 2% stratified audit sample (n = 287) of English-language comments is manually reviewed. This audit evaluates the plausibility of labels produced by the GPT- and BERT-based classifiers without assuming either as ground truth. Each label is assessed relative to the comment content, and summary metrics are calculated to reflect agreement with human judgment.

BERT yields an accuracy of 0.756, a macro-averaged F1-score of 0.548, and a Cohen's κ of 0.698. These values indicate high correspondence with human-judged plausibility. In contrast, the GPT-based model achieves lower performance on the same sample, with an accuracy of 0.557, macro-F1 of 0.360, and κ of 0.464. Performance metrics for both models, benchmarked against human plausibility, are depicted in Table 10.

model	*accuracy*	*macro_F1*	*kappa*
GPT	0.557	0.36	0.464
BERT	0.756	0.548	0.698

Table 10: Human Audit: Agreement Metrics for GPT and BERT

These audit results support the use of BERT as the preferred classifier for the English-language subset. They also provide an empirical basis for the harmonization protocol, in which BERT labels override GPT labels when available.

5.2.2 Agreement Between GPT and BERT Classifiers

This section reports on classifier agreement between the GPT- and BERT-based models across the full English-language subset. The comparison is based on aligned dominant emotion labels assigned by each classifier to the same input comment. GPT labels are used as a comparison baseline, without implying ground truth.

BERT achieves an overall accuracy of 0.436 relative to GPT labels, with a macro-averaged F1-score of 0.254 and a Cohen's κ of 0.319. Performance varies across emotion categories, with particularly low alignment among affectively adjacent labels such as admiration, approval, excitement, and joy. Additionally, comments misclassified as English by the GPT-based language filter are typically labeled as neutral by BERT. To visualize classifier divergence, a row-normalized confusion matrix is computed, showing the distribution of BERT predictions for each GPT label.

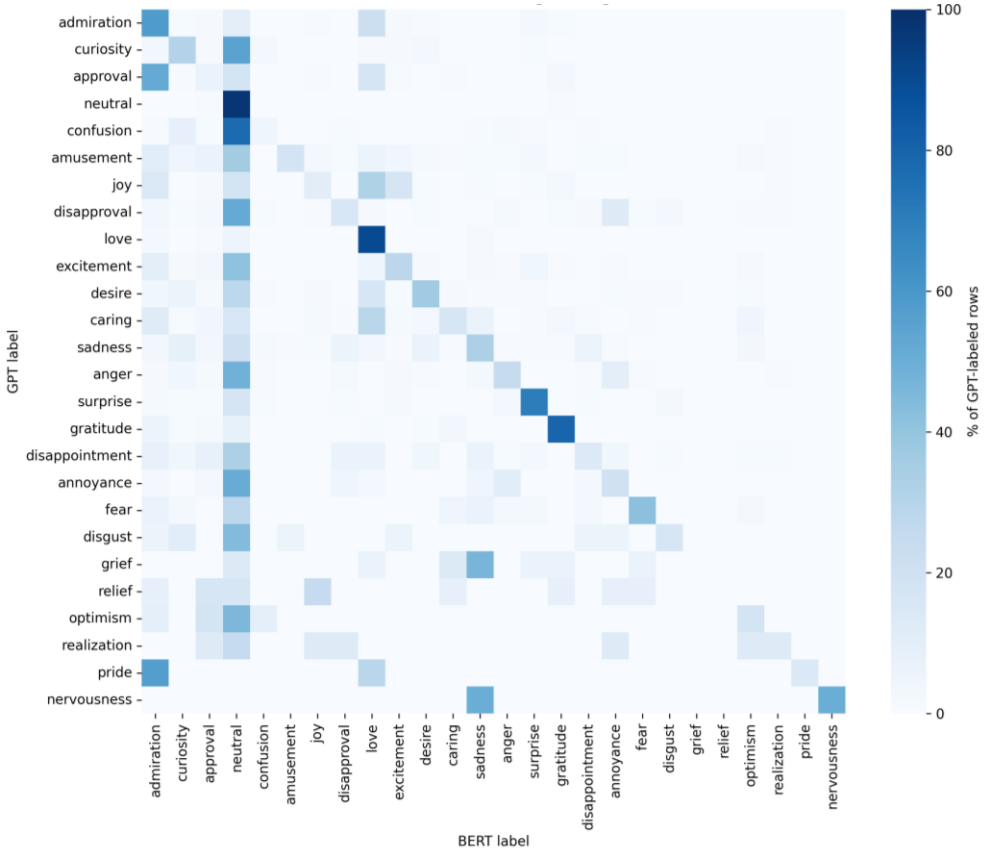

Figure 11: Confusion Matrix (GPT rows × BERT columns)

In addition to label-level agreement, classifier confidence is examined to assess model certainty. The average BERT confidence score across the English subset is 0.808, with notable variation by emotion label. GPT displays a slightly higher average confidence (0.848) but with tighter clustering across classes. Summarized distributions of classifier confidence are included in Appendix D for reference.

These findings support the harmonization procedure, in which BERT predictions are retained for English-language comments while GPT labels govern the remaining multilingual corpus.

5.2.3 *Consistency of Sentiment Polarity Derived from Emotion Labels*

This section evaluates the internal consistency of the classification pipeline by comparing sentiment polarity derived from GPT- and BERT-assigned emotion labels. Each emotion label is mapped to one of three sentiment polarity categories: Positive, Ambiguous, or Negative based on the GoEmotions taxonomy (Demszky et al. 2020). The goal of this procedure is to assess whether the two classifiers yield convergent affective interpretations at the sentiment level, despite discrepancies observed at the fine-grained emotion level.

The metrics indicate substantial polarity-level alignment between GPT and BERT emotion labels, with an overall accuracy of 0.811, a macro-averaged F1-score of 0.702, and a Cohen's κ of 0.628.

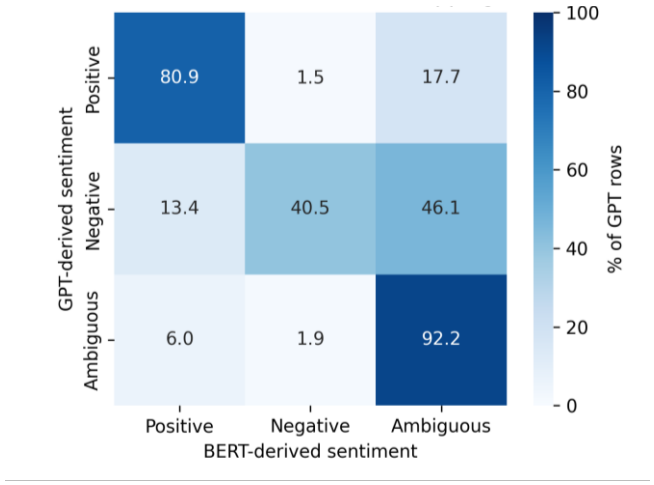

Figure 12: Confusion Matrix for Sentiment Polarity Agreement

These results substantiate the use of sentiment polarity as a complementary dimension to discrete emotion classification and reinforce its value to evaluate consistency within the annotation pipeline.

5.3 Emotional Response (H1)

This section reports results from a multinomial logit (MNL) model testing the effect of content origin (AI-generated vs. human-created) on the distribution of discrete emotional expressions in user comments. The model treats the dominant emotion label as the dependent variable and uses content origin as the sole predictor. The reference category is neutral, and effects are expressed as ORs relative to this baseline. Bootstrapped APPs for each emotion are also reported. Results are presented separately for the full multilingual dataset and the English-only subset.

5.3.1 *Model Summary and Screening Procedure*

Prior to modeling, rare emotion categories with fewer than 10 observations were removed (Embarrassment, Nervousness, Relief). A total of 64,789 observations were retained in the final model (English-only: 14,288). Both models converge and pass joint significance tests.

The pseudo-R^2 values indicate that the models explain a small portion of total variance, as is typical in text-based multinomial classification settings. However, the Wald χ^2 values confirm that the predictor variable (content origin) contributes significantly to the model fit in both samples.

	Full Sample Model	*English-Only Model*
N	64,789	14,288
Log-Likelihood (null)	-119,403	-27,564
Log-Likelihood (full)	-119,013	-27,222
McFadden pseudo-R^2	0.003	0.012
Wald χ^2, p	722.6, p < .001	655.0, p < .001

Table 11: Model Summary Statistics for Multinomial Logit Regression on Emotion Labels

5.3.2 Emotion-Specific Effects: Odds Ratios and Predicted Probabilities

This section presents the estimated effect of content origin on emotion labels in the full sample (N = 64,789), based on a multinomial logit model. The model uses the dominant predicted emotion as the outcome variable and compares the likelihood of each emotion occurring under AI-generated versus human-created content. The reference category is neutral.

Table 12 reports results for key emotion categories, displaying the odds ratio (OR) of each emotion under AI versus human origin, along with the corresponding average predicted probability difference (Δ(APP)). Emotions are ordered by OR, with values above 1 indicating an increased relative likelihood under AI origin, and values below 1 indicating a decreased likelihood. While ORs capture the relative shift in classification odds, Δ(APP) shows the absolute change in predicted expression rates. Full estimates, including non-significant effects, are provided in Appendix E.

Emotion	*OR*	*95% CI (OR)*	*p-value*	*Δ(APP)*
disapproval	3.22	[2.31–4.49]	<.001	+0.0043
amusement	1.94	[1.72–2.20]	<.001	+0.0173
joy	1.74	[1.55–1.94]	<.001	+0.0170
caring	1.70	[1.48–1.94]	<.001	+0.0101
annoyance	1.59	[1.11–2.29]	.012	+0.0010
anger	1.38	[1.01–1.89]	.038	+0.0010
sadness	1.34	[1.13–1.59]	.001	+0.0031
confusion	1.28	[1.10–1.49]	.001	+0.0033
love	1.12	[1.05–1.19]	<.001	+0.0181
surprise	0.82	[0.69–0.98]	.032	–0.0010
admiration	0.90	[0.85–0.95]	.001	–0.0490
curiosity	0.62	[0.55–0.70]	<.001	–0.0090
excitement	0.51	[0.44–0.59]	<.001	–0.0090
optimism	0.49	[0.31–0.78]	.003	–0.0009

Table 12: Emotion-specific effects of AI origin relative to human content

Notable shifts in the full dataset include increased predicted rates of love (+1.8pp), amusement (+1.7pp), and joy (+1.7pp), alongside a decrease in admiration (–4.9pp) and curiosity (–1.0pp) under AI origin.

5.3.3 Effects in the English Subset

This section re-estimates the multinomial logit model on the English-only subset (N = 14,288), using BERT-based emotion labels to assess whether the effect of content origin (AI vs. human) on emotion expression remains stable under high-confidence, monolingual conditions. Five rare categories with n < 10 were removed prior to modeling due to insufficient support (embarrassment, nervousness, relief, pride, grief).

Compared to the multilingual full dataset, the English-only model achieves a higher McFadden pseudo-R^2 (0.012 vs. 0.003), indicating that origin explains a slightly greater share of variance when language

is controlled. In both datasets, the omnibus Wald test rejects the null hypothesis of no content origin effect (Wald $\chi^2 = 655.0$, df = 22, $p < .001$).

Table 13 reports odds ratios (ORs) and average predicted probability differences (Δ(APP)) for emotion categories ordered by OR. Confidence intervals are shown for ORs. Δ(APP) values represent absolute shifts in predicted probabilities under AI versus human origin. Non-significant effects ($p > .05$) are indicated. Full estimates and Δ(APP) confidence intervals are provided in Appendix F.

Emotion	***OR***	***95% CI (OR)***	***p-value***	***Δ(APP)***
disapproval	2.37	[1.46–3.83]	<.001	+0.0090
annoyance	1.61	[1.04–2.49]	.032	+0.0069
fear	1.23	[0.57–2.63]	**.584**	+0.0015
anger	1.17	[0.75–1.82]	**.478**	+0.0042
caring	0.88	[0.54–1.45]	**.638**	+0.0019
amusement	0.81	[0.57–1.15]	**.246**	+0.0031
confusion	0.77	[0.50–1.19]	**.249**	+0.0017
joy	0.76	[0.54–1.09]	**.141**	+0.0025
sadness	0.75	[0.53–1.05]	**.096**	+0.0025
approval	0.73	[0.58–0.92]	.010	+0.0052
curiosity	0.59	[0.50–0.69]	<.001	+0.0009
excitement	0.50	[0.40–0.63]	<.001	–0.0030
optimism	0.49	[0.30–0.82]	.006	–0.0006
love	0.40	[0.36–0.45]	<.001	–0.0554
admiration	0.38	[0.34–0.41]	<.001	–0.1341

Table 13: Emotion-Specific Effects of AI Origin in the English Subset

In the English subset, admiration (–13.4pp) and love (–5.5pp) show the largest decreases in predicted probability under AI content. Disapproval (+0.9pp) and annoyance (+0.7pp) show the most pronounced increases.

5.3.4 Visual Comparison: Full Sample vs. English Subset

To enable direct comparison of emotion-level effects across both models, Figure 13 visualizes the average predicted probability differences (Δ(APP)) by emotion. Each bar represents the change in predicted probability of an emotion being assigned to a comment under AI-generated versus human-created content. Every emotion appears twice—once for the multilingual dataset and once for the English-only subset.

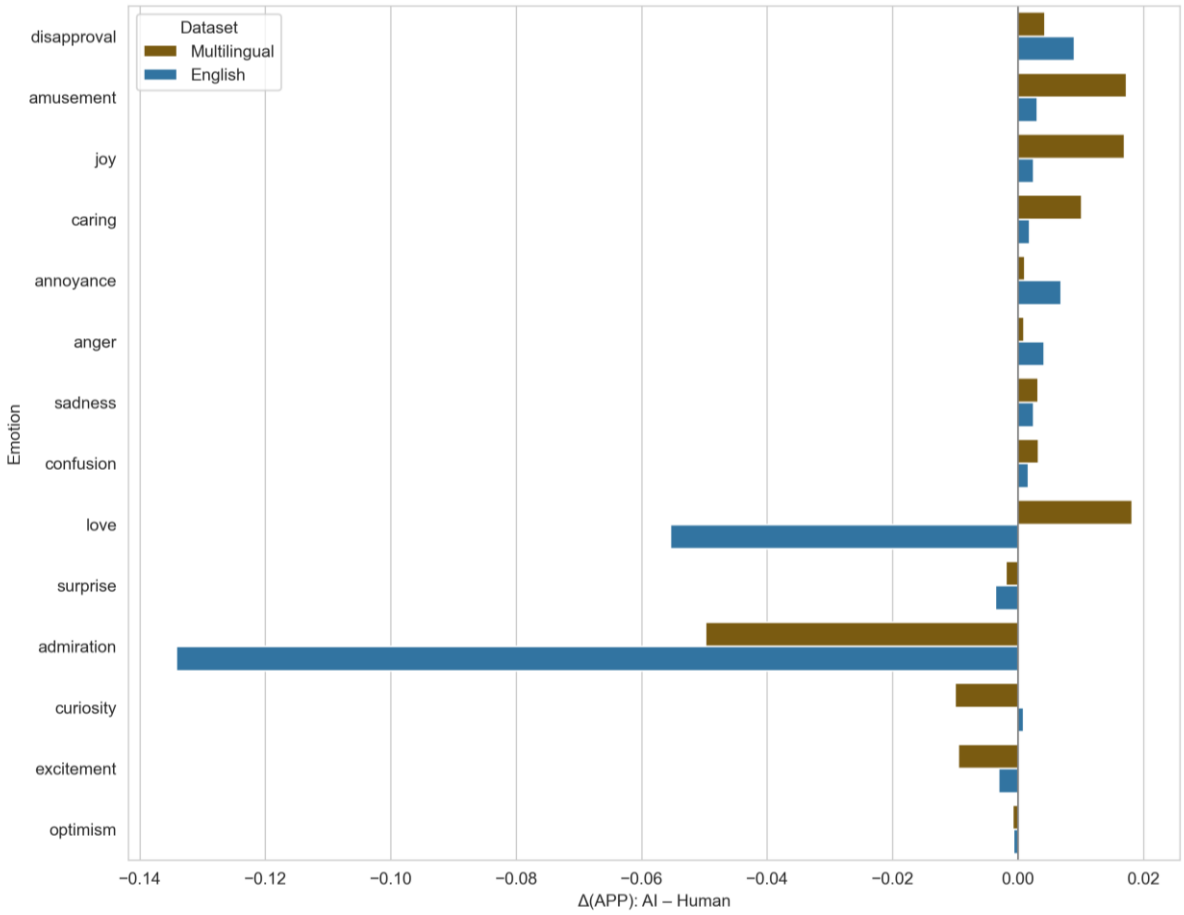

Figure 13: Predicted Probability Differences (Δ(APP)) by Emotion and Dataset

Figure 13 displays the predicted probability differences (Δ(APP)) for key emotions across the multilingual and English-only datasets. Most emotions show similar directional trends, but effect sizes vary.

Disapproval increases under AI origin in both datasets, with a stronger shift in the English subset. Amusement, joy, and caring show notable positive shifts in the full dataset, but these effects weaken in the English subset. Love shifts from a positive Δ(APP) in the full model to a negative value in English. Admiration shows a stronger decrease in the English subset than in the full dataset.

All Δ(APP) values shown are based on Table 12 and Table 13. Full estimates are reported in Appendix E and F.

5.4 Sentiment Polarity (H2)

This section reports the results for Hypothesis 2, which examines whether the valence of user sentiment differs depending on whether the content is AI-generated or human-created. Sentiment is classified using a three-level polarity scheme, Negative, Neutral, and Positive, assigned through GPT-based zero-shot annotation across the full multilingual dataset (N = 64,806). The analysis includes descriptive patterns, a statistical test of association, and examination of residuals and pairwise contrasts.

5.4.1 Distribution of Positive, Neutral, and Negative Labels

A descriptive overview of the sentiment distribution across the two content origin groups is shown in Table 14. While the absolute counts are higher for AI-generated content due to its larger share in the dataset, proportionally distinct patterns are visible.

Sentiment	*AI-Generated*	*Traditional art*
Negative	1,563	473
Neutral	5,619	2,401
Positive	36,359	18,391

Table 14: Frequency of sentiment categories by content origin (GPT-classified)

Visualizing these proportions, Figure 14 shows a stacked bar chart with row-normalized sentiment shares per origin group. Compared to human-created posts, AI-generated content is associated with a lower proportion of Positive sentiment, and higher proportions of Neutral and Negative sentiment.

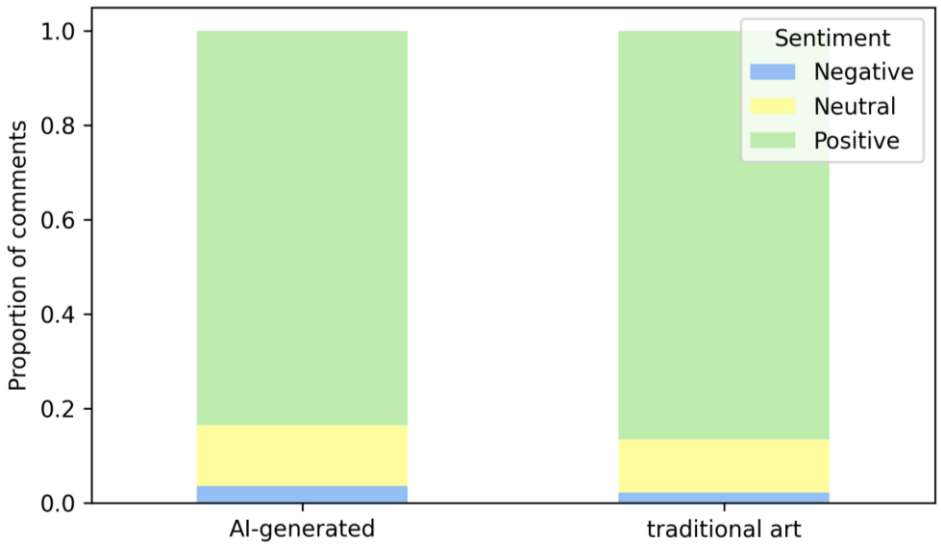

Figure 14: Sentiment distribution by content origin

5.4.2 Chi-Square Test and Effect Size

A chi-square test of independence is conducted to assess the relationship between content origin and sentiment distribution. The test yields a highly significant result: $\chi^2(2) = 129.88$, $p < .001$, indicating that the distribution of Negative, Neutral, and Positive sentiment labels differs systematically depending on whether the content is AI-generated or human-created.

The strength of this association, measured using Cramér's V, is 0.045, with a 95% confidence interval of [0.038, 0.053]. According to conventional benchmarks, this reflects a small but reliable effect size. While sentiment polarity is clearly associated with content origin, the effect accounts for only a limited portion of the total variation in responses.

5.4.3 Standardized Residuals and Pairwise Differences

To understand which sentiment categories contribute most strongly to the observed association, standardized residuals are inspected. These residuals show the degree to which each observed value deviates from the expected value under a model of independence. Table 15 summarizes the results.

Origin	*Negative*	*Neutral*	*Positive*
AI-generated	+5.27	+3.14	–2.22
Traditional art	–7.55	–4.50	+3.18

Table 15: Standardized residuals from the contingency table

These values indicate that AI-generated content received more Negative and Neutral sentiment labels than expected, while Positive labels were less frequent than expected. The opposite holds for human-

created content, which received more Positive and fewer Negative or Neutral sentiment labels than would be expected if origin and sentiment were unrelated.

The heatmap in Figure 15 visualizes these standardized residuals by cell, highlighting the strength and direction of the deviation.

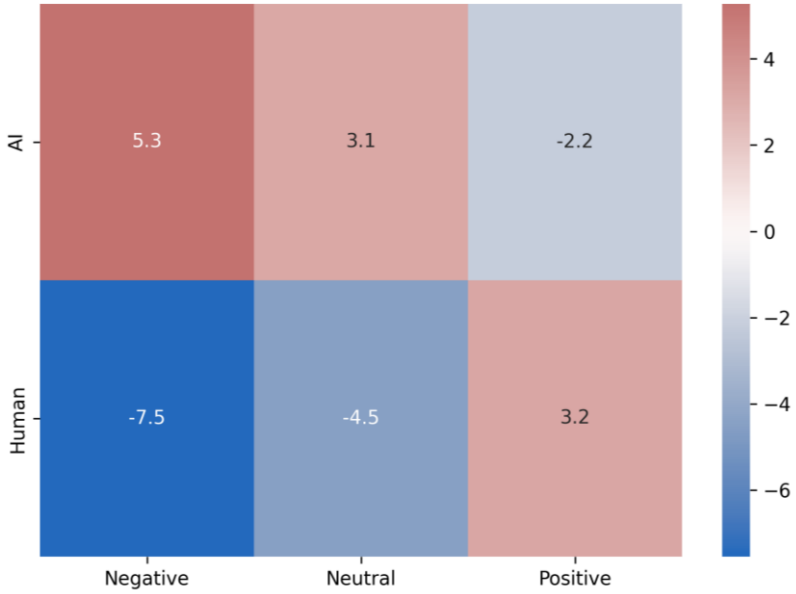

Figure 15: Standardized residuals for Origin × Sentiment

To further dissect the sentiment pattern, pairwise chi-square tests are conducted between all three sentiment categories. After applying Holm correction for multiple comparisons, all contrasts remain highly significant, indicating that the two origin groups differ not only in the overall distribution but in each pairwise combination of sentiment.

Contrast	χ^2	*p (Holm-corrected)*
Negative vs. Neutral	35.44	2.6×10^{-9}
Negative vs. Positive	94.49	7.4×10^{-22}
Neutral vs. Positive	41.98	1.8×10^{-10}

Table 16: Pairwise chi-square contrasts between sentiment categories

Together, these findings confirm that sentiment polarity is significantly associated with content origin, with AI-generated content more likely to elicit Neutral and Negative sentiment, and human-created content more likely to elicit Positive sentiment. While the overall effect is small in size, it is statistically robust and consistent across comparisons.

5.5 Behavioral Engagement (H3)

This section evaluates whether content origin (AI-generated vs. human-created) predicts differences in user engagement behavior on social media. Engagement is operationalized through three observable metrics: likes, comments, and reshares. Each metric is modeled independently using regression techniques appropriate to its distribution, with model-based predicted means and confidence intervals reported alongside effect sizes.

5.5.1 Like Count Analysis

Of 5,001 posts, the top 1% outliers (≥41,189 likes) were excluded, yielding a final sample of 4,951. Likes remain non-zero for all posts. The median number of likes is 582, and the mean is 4,595 (SD = 125,748), reflecting a long-tailed distribution influenced by highly visible content.

A Negative Binomial regression model is used to estimate the relationship between content origin and like count. The model reveals substantial overdispersion ($\hat{\alpha} = 2.17$), justifying the use of a count model beyond the Poisson framework. The coefficient for content origin is statistically significant ($\beta = 0.153$, SE = 0.043, $z = 3.57$, $p < .001$), with an incidence-rate ratio (IRR) of 1.165 (95% CI [1.071, 1.268]). This result indicates that AI-generated posts receive approximately 16.5% more likes than human-created posts, all else equal.

Origin	*Predicted Mean*	*95% CI*
Traditional art	1,399	[1,310, 1,494]
AI-generated	1,630	[1,547, 1,718]

Table 17: Predicted Mean Likes by Content Origin

5.5.2 Comment Count Analysis

All 5,001 posts are retained for the comment count analysis, as the distribution, while skewed, is not dominated by outliers. The median number of comments is 14, with a mean of 54.8 (SD = 690.9). Approximately 14.5% of posts have zero comments, which does not warrant the use of a zero-inflated model.

The Negative Binomial regression model again shows substantial overdispersion ($\hat{\alpha} = 2.10$). The origin coefficient is large and statistically significant ($\beta = 0.486$, SE = 0.043, $z = 11.41$, $p < .0001$), yielding an incidence-rate ratio (IRR) of 1.626 (95% CI [1.496, 1.768]). This implies that, on average, AI-generated posts attract 62.6% more comments than human-created ones.

Origin	*Predicted Mean*	*95% CI*
Traditional art	23.44	[21.95, 25.02]
AI-generated	38.12	[36.19, 40.15]

Table 18: Predicted Mean Comments by Content Origin

5.5.3 Reshare Count Analysis

Reshare counts are markedly zero-inflated, with 64.6% of posts receiving no reshares at all. The median is zero, and the mean is 1,706 (SD = 54,101). The top 1% of posts (n = 50), each with 7,588 or more reshares, are removed to reduce extreme skew, resulting in a modeling sample of 4,951 posts.

Given the large proportion of zeros, a Zero-Inflated Negative Binomial model is used. The model converges successfully (LL = –13,893; LLR $p = 2.28 \times 10^{-12}$), and both the inflation and count components are significant. The inflation term (inflate_const = 0.142, SE = 0.069, $p = .040$) suggests that approximately 53.5% of posts belong to a structural-zero group. In the count model, content origin is a significant predictor ($\beta = 0.739$, SE = 0.101, $z = 7.35$, $p < .001$), with an incidence-rate ratio (IRR) of 2.094 (95% CI [1.719, 2.550]). This corresponds to an expected increase of 109% in reshare count for AI-generated posts.

Origin	*Predicted Mean*	*95% CI*
Traditional art	65.15	[61.05, 69.53]
AI-generated	149.70	[142.16, 157.64]

Table 19: Predicted Mean Reshares by Content Origin

5.5.4 Engagement Summary

To enable direct comparison across all engagement metrics, presents the incidence-rate ratios for likes, comments, and reshares, along with their 95% confidence intervals. The strongest effect is observed for reshares, followed by comments, then likes. In this context, the IRR expresses the relative change in expected engagement volume for AI-generated posts compared to human-created posts.

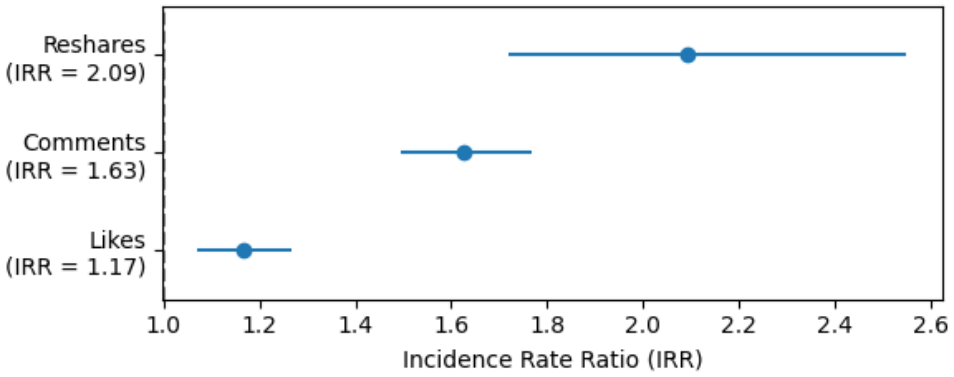

Figure 16: Incidence-rate ratios (IRR) for likes, comments, and reshares

A second comparative view is provided in Figure 17, which shows model-based predicted engagement volumes for likes, comments, and reshares, displayed separately for AI-generated and traditional art. Error bars represent 95% confidence intervals.

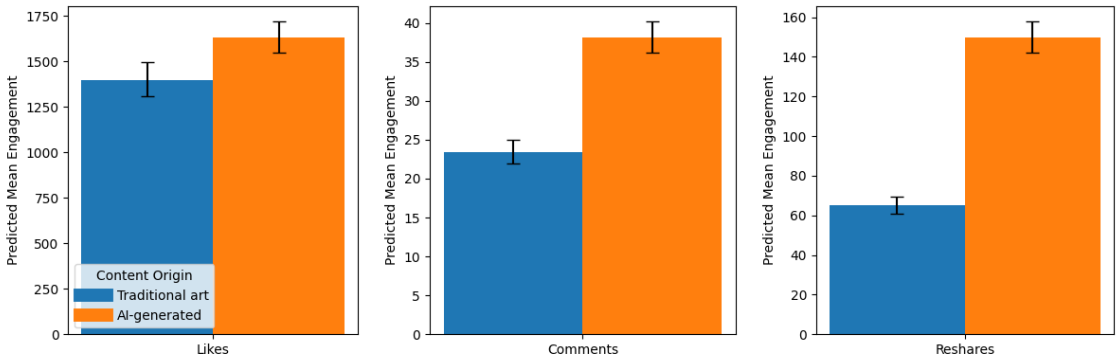

Figure 17: Model-based predicted engagement volumes by content origin

5.6 Depth of Engagement (H4)

This section presents results for Hypothesis 4, which examines whether the depth of user engagement differs systematically between AI-generated and human-created art. Engagement depth is categorized into four ordered levels, Superficial, Moderate, Deep, and Very Deep, based on qualitative indicators in user comments. The analysis includes a contingency table, standardized residuals, post hoc comparisons, and a multinomial logit model to quantify the association.

5.6.1 *Descriptive Distribution and Contingency Analysis*

Table 20 presents the raw frequency of comments by engagement depth and content origin. A total of N=64,806 comments were analyzed.

Engagement Depth	*AI-Generated*	*Traditional Art*
Superficial	37,301	14,949
Moderate	5,532	5,530
Deep	631	713
Very Deep	77	73

Table 20: Frequency of engagement depth levels by content origin

A chi-square test of independence reveals a highly significant association between content origin and engagement depth: $\chi^2(3) = 2166.0$, $p < .001$. The strength of this association, measured using Cramér's V, is 0.183 (95% CI [0.175, 0.191]), reflecting a small to moderate effect size.

Figure 18 visualizes the relative distribution of engagement depth levels, row-normalized within each origin group. A pronounced concentration of Superficial responses is observed for AI-generated content, whereas deeper levels are more evenly distributed under human-created posts.

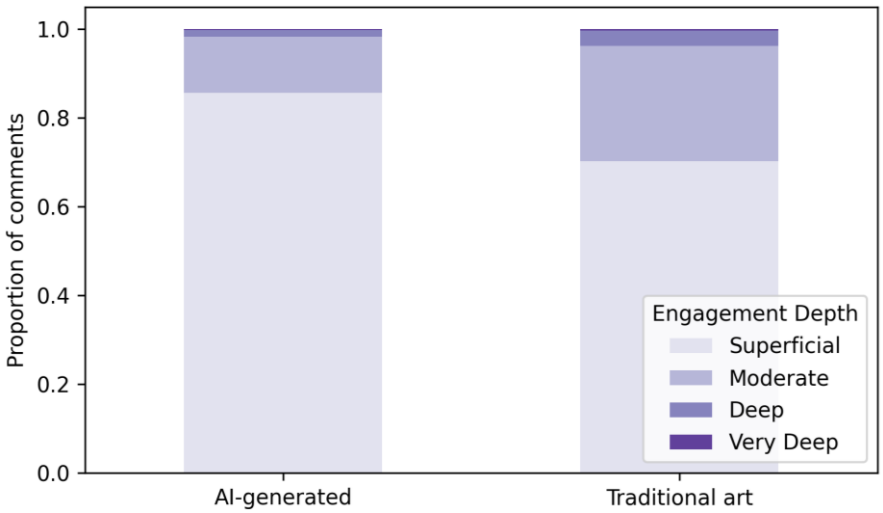

Figure 18: Engagement depth distribution by content origin

5.6.2 *Standardized Residuals and Pairwise Comparison*

To identify which categories contribute most to the overall association, standardized residuals are examined (Table 21). These residuals indicate that AI-generated posts are significantly overrepresented in the Superficial category and underrepresented in all deeper levels. Human-created content shows the opposite pattern.

Origin	*Superficial*	*Moderate*	*Deep*	*Very Deep*
AI-generated	+11.72	–22.04	–9.05	–2.37
Traditional art	–16.77	+31.54	+12.95	+3.39

Table 21: Standardized residuals for engagement depth

Residuals confirm that AI content receives disproportionately more superficial comments, with residuals exceeding ±10. Holm-corrected pairwise chi-square tests confirm that Superficial engagement differs significantly from each deeper level (all $p < .001$). No significant differences are observed between Moderate, Deep, and Very Deep levels (see Appendix G).

5.6.3 Multinomial Logit Model

To quantify the relationship between content origin and engagement depth in a model-based framework, a multinomial logit model is estimated using Superficial engagement as the reference category. The model includes a binary origin predictor (AI vs. Human) and converges successfully, yielding a pseudo R^2 of 0.028 and a highly significant model fit (Log-Likelihood = –35,895; LL-Null = –36,928; LLR $p < .001$). All origin coefficients for the deeper engagement categories are statistically significant at $p < .001$.

Engagement Depth	*OR (AI vs. Human)*	*95% CI*
Moderate	0.401	[0.384, 0.418]
Deep	0.355	[0.318, 0.395]
Very Deep	0.423	[0.307, 0.583]

Table 22: Odds ratios from multinomial logit model (reference = Superficial)

Odds ratios below 1 indicate lower odds of a given engagement level compared to Superficial, for AI content relative to human content. The odds ratios indicate that AI-generated posts are substantially less likely to elicit Moderate, Deep, or Very Deep engagement compared to human-created posts. The strongest attenuation is observed for Deep engagement, followed closely by Moderate.

5.7 Moderating Role of Exposure Pattern (H5)

This chapter presents the results for Hypothesis 5, which evaluates whether the emotional effect of content origin is moderated by users' exposure pattern. Specifically, it tests whether users who interact with both AI-generated and traditional artworks ("Mixed") respond differently than those exposed to only one content type ("AI-only" or "Traditional-only"). The analysis is implemented using a regularized multinomial logit model with interaction terms, supplemented by predicted probabilities.

5.7.1 Interaction Effects by Exposure Group

Users are categorized into three exposure groups: 28,254 users are classified as AI-only, 17,618 as Traditional-only, and 136 as Mixed exposure. To enable stable estimation, emotion categories absent in any exposure–origin combination are excluded. The final model includes twelve emotions, with neutral as the reference category.

A ridge-penalized multinomial logit model is estimated using content origin, exposure group, and their interaction as predictors. An omnibus bootstrap-based Wald test for interaction yields $\chi^2(22) = 725.26$, $p = .732$, indicating no significant moderation by exposure group.

Despite the non-significant test, interaction terms are reported descriptively in Table 23. No odds ratio differs significantly from baseline after Holm correction (all adjusted $p > .05$). While some directional differences appear, for example Love in the Mixed group (OR = 1.69) and Admiration (OR = 1.36), confidence intervals are wide and effects remain statistically inconclusive.

Emotion	*OR (AI-only)*	*95% CI*	*OR (Mixed)*	*95% CI*
Admiration	0.85	[0.79, 0.91]	1.36	[1.13, 1.68]
Amusement	1.33	[1.24, 1.44]	0.77	[0.63, 0.92]
Anger	1.12	[1.02, 1.26]	0.92	[0.88, 0.93]
Joy	1.17	[1.04, 1.30]	1.01	[0.74, 1.42]
Love	0.85	[0.78, 0.91]	1.69	[1.37, 2.21]

Table 23: Interaction effect sizes for AI-only and Mixed groups (ORs for Origin × Exposure)

5.7.2 Emotional Probabilities Across Exposure Groups

Average predicted probabilities (APPs) are computed across all six combinations of content origin and user exposure. The resulting profiles do not display systematic or substantial divergence. While some emotion categories show numeric differences, for example, slightly higher admiration under AI content for Mixed users, these patterns fall within the margin of estimation error. Love and joy show elevated values across groups, but again without interaction-level significance.

Taken together, the results suggest that users' prior engagement with AI or human content does not meaningfully moderate their emotional responses to a given post. Content origin remains the primary driver, independent of exposure group.

5.8 Summary of Findings

Content origin consistently shapes user response across affective and behavioral dimensions. Emotion classification reveals that AI-generated posts elicit more disapproval, while traditional art evokes greater admiration, curiosity, and excitement. These differences remain stable across both multilingual and English-only datasets.

Sentiment polarity follows a similar pattern. AI content is associated with higher rates of neutral and negative sentiment, and lower rates of positive sentiment, relative to human-created art. Although the effect size is modest, the direction is statistically robust.

Behavioral engagement is higher for AI-generated content across all three metrics. Posts tagged as AI receive more likes, attract more comments, and are reshared more frequently, with the strongest increase observed for reshares.

Engagement depth diverges notably by origin. Comments on AI-generated content cluster around superficial engagement, while deeper responses are more frequent under traditional art. These origin effects are statistically significant across all depth levels.

User exposure history does not moderate emotional responses. Interaction models show no systematic difference between users engaging with both content types and those exposed to only one. The effect of origin persists independently of prior interaction patterns.

6 Discussion

This chapter addresses the thesis' central research question - How do users' emotional responses to explicitly tagged AI-generated content compare with their responses to human-created content on social media? - and situates the empirical findings within the wider scholarly discourse. The results indicate a clear pattern: when Instagram posts are openly labeled as AI-generated, users exhibit more negative evaluative emotions (most notably a decline in admiration and a rise in disapproval), engage in significantly higher but markedly shallower interaction, and offer fewer deep, reflective comments than they do for comparable human-created art.

The remainder of the chapter articulates the theoretical implications of these findings, derives practical and policy lessons, acknowledges limitations, and outlines directions for future research.

6.1 Implications for Theory

6.1.1 Transparency Cues as Signals and Affective Triggers

Labeling AI-generated content on Instagram produces a dual emotional pattern. Admiration declines significantly (–4.9 percentage points; OR = 0.90), while disapproval increases more than threefold (+0.4 pp; OR = 3.22). At the same time, small but consistent increases are observed in amusement, joy, and caring (see Table 12).

This combination reflects a cognitive-affective tension. Drawing on Signaling Theory (Spence 1973), the "#aiart" tag operates as a heuristic that foregrounds machine authorship. As a result, users appear to discount the underlying effort and authenticity of the content, an interpretation consistent with the observed decrease in admiration. The fact that this penalty emerges in a high-noise, real-world platform setting reinforces earlier experimental findings (Bellaiche et al. 2023; Gabbiadini et al. 2024), while also extending their external validity into dynamic social environments (Bauer et al. 2024).

In parallel, the novelty of algorithmic authorship evokes mild exploratory emotions. According to Arousal-Curiosity Theory (Berlyne 1960), novel or unexpected stimuli tend to stimulate attentional and affective responses such as amusement even in the absence of endorsement. This explains the modest increases in light positive affect despite a broader cooling of evaluative responses. Applying this theory further, the aversive tension displayed in the data may be partially explained by the very high novelty of GenAI content.

The observed rise in disapproval further supports the symbolic-threat framing proposed by Gabbiadini et al. (2024), where AI-generated content is interpreted as encroaching on domains traditionally reserved for human creativity. Although direct threat-related emotions such as fear remain rare, possibly due to the performative and curated nature of Instagram, disapproval may serve as a socially palatable proxy. The response pattern is also consistent with the Uncanny Valley hypothesis (Mori et al. 2012), which suggests that discomfort and disapproval increase when artifacts closely but imperfectly resemble human attributes such as creativity, thus triggering aversive emotional reactions.

Viewed through the lens of algorithm aversion (Dietvorst et al. 2014), provenance cues also act as cognitive shortcuts that highlight machine fallibility and the absence of human intentionality. This affective signal, in turn, may serve as an early-stage precursor to deeper skepticism. While the present data do not directly measure long-term trust trajectories, the emotional shift observed here, especially the decline in admiration and rise in disapproval, echoes the initial conditions for downstream distrust noted in longitudinal accounts of algorithmic trust formation (Lukyanenko et al. 2022).

In sum, labeling practices shape more than just awareness. They operate as affective and cognitive filters, initiating interpretive shifts that reframe the artistic status of the work and potentially anchor user attitudes toward generative systems.

6.1.2 Volume Without Depth: An Engagement Paradox

Despite prompting more negative or ambivalent emotional responses, AI-generated content elicits significantly higher levels of visible interaction: Likes increase by 16%, Comments by 63%, and Reshares by 109% (IRRs > 1.15, $p < .001$). Yet this surface-level popularity does not translate into deeper engagement. The odds of receiving a Deep or Very Deep comment are reduced by approximately 60-65%, revealing a striking gap between attention and elaboration.

This asymmetry raises challenges for established engagement frameworks. Models such as the Customer Engagement Model (Brodie et al. 2011) and the COBRA hierarchy (Schivinski et al. 2016) treat contribution behaviors (likes, shares, comments) as indicative of active user involvement. However, the findings here show that AI-generated art drives high-frequency participation without the sustained reflection or emotional investment that these metrics may imply. In this sense, the results extend prior models by exposing a structural disconnect between engagement volume and engagement quality in algorithmically mediated creative contexts.

From the perspective of Uses and Gratifications Theory (Katz et al. 1973; O'Day and Heimberg 2021), this pattern reflects selective gratification. AI art appears well suited to satisfying low-effort psychological needs, including entertainment and social signaling, due in part to its novelty and frictionless shareability. However, it less frequently supports more cognitively or emotionally demanding gratifications such as identity expression, aesthetic immersion, or epistemic curiosity, which are more typical of traditional art engagement.

This differentiation aligns with Verduyn et al. (2017) observation that high-frequency, low-depth interaction is often decoupled from psychosocial benefit. It also suggests that behavioral engagement metrics, while convenient, may overstate the user's psychological or emotional connection to content, especially when algorithmic novelty inflates visibility but not value.

6.1.3 Cultural Appraisal and Emotional Amplification

In the English-only subset, the emotional asymmetry sharpens. Admiration declines by 13.4 percentage points and love by 5.5 points, while disapproval and annoyance rise further. This pattern reflects a combination of methodological and cultural dynamics.

Methodologically, BERT refinement removes emoji-dominant praise that was previously classified as positive by GPT, exposing a more critically worded textual layer. Culturally, English-speaking users operate within low-context, high-individualism cultures (Hofstede 2011) that favor explicit judgment and direct emotional expression (Dewaele 2010). These cultural norms are further shaped by the elevated prominence of AI ethics discourse in Anglophone media (Corrêa et al. 2023), which primes audiences to interpret AI-generated content in terms of authenticity risks and existential threat.

From the lens of Appraisal Theory (Smith and Lazarus 1993), these contextual frames lower the activation threshold for symbolic-threat appraisals (Gabbiadini et al. 2024). When provenance is disclosed, English-speaking users may be more likely to interpret the work as misaligned with their expectations for effort, creativity, or authorship, thereby intensifying disapproval. In this light, cultural positioning modulates not only emotional expression, but the underlying interpretive gate through which AI-generated content is evaluated.

This finding complicates the notion of universal algorithm aversion. Rather than assuming homogenous emotional responses across audiences, it points to the need for future research to treat cultural context as a moderator, especially in cross-lingual affective computing. Language is not merely a classification artifact, but a proxy for differing appraisal schemas.

6.1.4 Familiarity Without Reappraisal

Hypothesis 5, which tests whether prior exposure to both content types moderates emotional response, is not supported. Users who comment on both #aiart and #traditionalart posts (Mixed group, n = 136) show no statistically distinct emotion profile (interaction $\chi^2(22) = 725$, $p = .73$).

This absence of differentiation may partly stem from statistical underpowering. The small size of the Mixed group suggests that traditional-art consumers may systematically avoid #aiart content, reducing the likelihood of meaningful cross-exposure and limiting contact with its stylistic or conceptual cues.

Moreover, from the perspective of Conceptual Act Theory (Barrett 2006), once a category such as “AI art” becomes emotionally associated with skepticism or symbolic threat, mere exposure may be insufficient to revise that schema, particularly if deeper cultural or identity-related tensions remain unresolved.

6.1.5 Summary: Affective Structures and Methodological Insights

The five hypotheses collectively reveal a consistent affective and behavioral pattern. Transparent disclosure of AI authorship reframes creative content from a signal of artistic mastery to a technological artifact. This shift increases surface-level interaction (H3) but dampens reflective engagement (H4) and reduces positive evaluative emotions such as admiration (H1) and overall sentiment positivity (H2).

While light exploratory emotions such as amusement, joy, and caring increase modestly, more cognitively engaging emotions like curiosity and excitement decline, indicating that user attention is captured but not deeply activated. These shifts coincide with a rise in disapproval, reinforcing the symbolic-threat interpretation.

The result is a dual-layered response: visible engagement without interpretive investment, emotional reaction without depth. This pattern spans emotional valence, engagement type, and user exposure history (H5). It suggests that provenance cues do not merely inform but actively shape users’ affective and cognitive orientation toward content. In this way, transparency operates as a perceptual regulator within algorithmically mediated social environments. Speculatively, this initial novelty-driven engagement and amusement might attenuate over time, potentially shifting towards boredom or even contempt, as repeated exposure reduces novelty and reinforces underlying negative appraisals, a dynamic warranting longitudinal investigation.

6.2 Implications for Practice and Policy

The empirical findings presented in this study have substantial implications for practice and policy, particularly regarding the management and disclosure of AI-generated content on digital platforms. Central to these implications is the recognition that transparency mechanisms do more than fulfill regulatory obligations. They actively shape user perception, emotional engagement, and interpretive frameworks. The observed emotional responses, notably increased disapproval and diminished admiration toward AI-labeled content, suggest that provenance cues significantly influence the perceived authenticity and creative legitimacy of digital artifacts. Platforms thus have ethical and practical responsibilities not only to disclose the origins of AI-generated content but also to thoughtfully manage the nuanced effects these disclosures entail (Larsson and Heintz 2020; Lund et al. 2025).

Practically, these results highlight the need for transparency strategies that are context-sensitive rather than uniformly applied. Static disclosures, such as the simple “AI-generated” labels, risk reinforcing negative emotional responses and amplifying algorithm aversion (Dietvorst et al. 2014; Park et al. 2024). Instead, flexible and layered transparency mechanisms, such as embedded metadata standards exemplified by the C2PA standard (C2PA 2024), may be used to better balance regulatory requirements with user experience. These mechanisms could allow users to engage with provenance information in ways proportionate to their interest or informational needs, thereby mitigating the potential negative impacts associated with overt foregrounding of AI authorship.

Furthermore, the observed asymmetry between high-volume but low-depth engagement with AI-generated content suggests that platforms and content creators should reconsider traditional engagement metrics (Schivinski et al. 2016; Trunfio and Rossi 2021). High levels of superficial engagement such as likes, comments, and reshares, may not accurately reflect deeper emotional or cognitive investment. This aligns with findings from Verduyn et al. (2017), who caution against interpreting high-frequency, low-depth interactions as indicative of meaningful psychological connections. Therefore, practitioners should adopt a more nuanced understanding of engagement, emphasizing qualitative dimensions such as emotional resonance, perceived authenticity, and user trust.

These findings also carry implications for content creators, who may benefit from a more refined understanding of audience responses to AI-generated versus traditionally created content. With current transparency regulations shaping emotional reactions, creators can strategically decide when and how to leverage generative tools, weighing novelty and productivity gains (Doshi and Hauser 2023; Zhou and Lee 2024) against potential declines in perceived authenticity and deeper emotional engagement. Likewise, brands and companies should consider how provenance disclosure affects their public communication strategies, possibly favoring traditionally created media to cultivate longer-lasting emotional relationships with customers and avoid triggering inherent biases or negative associations (Gillath et al. 2021; Lukyanenko et al. 2022).

Moreover, the multilingual variations identified, particularly the amplified negative emotional responses within English-speaking contexts highlight the importance of culturally sensitive policy frameworks. Transparency measures designed under monolithic assumptions risk failing to accommodate diverse interpretive frameworks and emotional appraisal mechanisms across cultures (Dewaele 2010; Hofstede 2011). Consequently, culturally nuanced approaches to AI-content disclosure, potentially including differentiated or localized transparency guidelines, could better align policy standards with varying user expectations and emotional norms.

In summary, effective ethical stewardship of AI-generated content extends beyond mere regulatory compliance. Adopting disclosure practices that respect user autonomy, accommodate interpretive flexibility, and remain responsive to evolving human-AI interactions will help ensure that transparency enhances rather than undermines user trust, engagement depth, and perceived authenticity in digitally mediated creative contexts.

6.3 Study Limitations and Boundary Conditions

The empirical findings presented in this thesis offer robust and externally valid evidence regarding the influence of Content Origin (AI-generated vs. human-created) on user emotions and engagement behaviors. However, these findings must be interpreted within several methodological, analytical, and ethical limitations. Clearly articulating these boundaries not only prevents overgeneralization but also directs future research towards addressing specific gaps and uncertainties.

6.3.1 Sampling Frame and Platform Ecology

Data collection focused exclusively on Instagram's algorithmically-curated "Top" posts for hashtags #aiart and #traditionalart within a single month (8 Mar – 8 Apr 2025). As a result, findings reflect user reactions to already highly visible content, potentially inflating baseline engagement metrics. Generalizability to less prominent content, niche communities, or other platforms like text-centric spaces or specialist art communities remains uncertain. Future research should explicitly compare across platforms and engagement visibility levels to test these boundaries.

6.3.2 Cross-Sectional Snapshot

The analysis captures user responses during a period of high salience and novelty of GenAI art, providing no insight into long-term changes in user sentiment or habituation effects. Transparency-induced attitudes may intensify or attenuate over time as AI-generated content becomes more common or perceptions shift. Therefore, these findings should be considered initial evidence of a potentially dynamic phenomenon rather than a stable long-term trend. Future longitudinal or panel designs could explicitly measure temporal shifts in user attitudes toward AI-generated content.

6.3.3 Hashtag-Based Origin Coding and Potential Misclassification

Content provenance was inferred solely from creator-supplied hashtags (#aiart or #traditionalart). Mis-tagging or strategic hashtag use may have introduced false positives or negatives. Although the robustness of observed effects suggests real provenance-driven differences, automated methods such as future integration of C2PA detection standards could significantly improve attribution accuracy.

6.3.4 Emotion Labeling Accuracy and Taxonomic Limitations

The dual-stage GPT–BERT classification pipeline reaches moderate accuracy and thus includes inherent labeling errors. The GoEmotions dataset is itself Reddit-derived and under-represents multilingual, non-Western dialects, and emoji-heavy communication (Demszky et al. 2020), potentially limiting the accuracy of emotional classifications for a global user base. The use of single dominant emotions per comment further simplifies nuanced affective blends, likely biasing subtle emotional responses toward neutrality.

6.3.5 Nested Data and Unmodelled Dependence

Comments analyzed are nested within posts and users, creating inherent statistical dependencies. Treating comments as independent observations likely underestimates standard errors, inflating the significance of findings. Future work employing hierarchical or multilevel models could more precisely model and estimate intra-class correlations, improving inference accuracy.

6.3.6 Cultural and Linguistic Generalizability

Despite covering 62 languages, the refinement of emotion classification via BERT is limited exclusively to English-language comments. This introduces differential classifier accuracy and neglects cultural variability in emotional expression norms. Consequently, multilingual results represent directional averages rather than culturally invariant parameters. Explicit cross-cultural moderation studies are recommended to clarify cultural boundary conditions.

6.3.7 Ethical Constraints and Technical Risks

Ethically compliant pseudonymisation, including salted SHA-256 hashing and a twelve-month data retention policy, protects user identity yet precludes the use of potentially insightful metadata. Additionally, reliance on third-party scraping rather than official Instagram APIs poses a compliance and

replicability risk, as platform policies and technical availability may change unexpectedly. Researchers must navigate these ethical and technical trade-offs carefully in future studies.

6.3.8 Statistical Modeling Constraints

Several analytical decisions impose additional interpretive constraints. Collapsing or removing low-frequency emotion categories due to data sparsity may bias results or obscure subtle emotional effects. Multiple simultaneous statistical tests introduce the risk of inflated Type I errors, partially but not entirely mitigated by Holm corrections. Distributional assumptions inherent in Negative Binomial Regression (e.g., constant dispersion parameters) may not fully reflect highly variable engagement patterns, particularly when posts experience viral spikes or structural zeroes.

6.3.9 External Validity and Platform-Specific Interpretive Scope

Given the Instagram-specific context of this study, caution is warranted in extending findings to platforms with differing affordances, cultural norms, and user demographics. Behavioral metrics, while objective, capture only publicly visible engagement, omitting private user interactions, viewing durations, and platform-specific emotional norms.

6.3.10 Subjectivity in Human Audit

A methodological limitation lies in the human audit process used to validate emotion and engagement classification outputs. In this study, audits were conducted by the researchers themselves, which may introduce subjective bias and inadvertently reinforce prior assumptions. While domain familiarity aids interpretive accuracy, it also increases the risk of confirmation bias. Employing external or blinded auditors in future studies would enhance classification reliability, offer a more impartial validation layer, and improve the generalizability of interpretive insights.

6.3.11 Authenticity of User Accounts

A final limitation concerns the assumption that all commenting entities represent genuine human users. Due to the absence of demographic or behavioral metadata, it remains uncertain whether some of the accounts interacting with either AI-generated or traditional content are automated agents, bots, or hybrid algorithmic accounts. While the study treats each comment as a discrete, human-authored signal, this assumption cannot be empirically verified within the current dataset. Future research should aim to incorporate metadata or authentication heuristics to distinguish authentic user activity from automated amplification and assess its influence on emotion and engagement metrics.

6.4 Future Research Directions

The empirical and methodological boundaries identified in this research open multiple avenues for advancing knowledge on algorithmic authorship and user responses to AI-generated content. Building upon these insights, several clear research priorities emerge, each designed to address specific gaps or refine existing findings.

Firstly, adopting a longitudinal research design would significantly enhance understanding of the temporal dynamics of user responses. While this study provides robust initial evidence, it cannot capture changes in emotional or cognitive responses over time. Future research should employ panel studies to assess whether negative initial reactions (e.g., diminished admiration or increased disapproval) persist, intensify, or attenuate as users habituate to AI-generated content. Such longitudinal insights would contribute valuable evidence on whether algorithm aversion is a transient novelty effect or a stable psychological response.

Secondly, cross-platform comparisons could significantly improve ecological validity. The current findings are platform-specific, emerging from Instagram's visually-centric and algorithmically curated environment. Subsequent studies should replicate these analyses across different digital platforms, including text-heavy environments like Reddit or Twitter, short-video platforms like TikTok, and specialized art communities such as DeviantArt. Examining different media ecosystems will clarify how platform affordances and community norms influence both the breadth and depth of engagement with AI-generated content.

Thirdly, cultural and linguistic variability demands more rigorous and explicit exploration. The observed amplification of negative reactions among English-speaking users suggests cultural moderation. To isolate linguistic classification biases from genuine cultural differences in emotional appraisal, future research should systematically sample diverse linguistic and regional populations. Moreover, culturally sensitive classification pipelines, possibly fine-tuned on region-specific datasets, would ensure more precise and reliable emotion labeling, improving theoretical clarity on symbolic-threat appraisals across cultures.

Fourthly, the granularity and format of provenance disclosures should be explored experimentally. Current binary disclosures (e.g., "AI-generated") may oversimplify provenance information, unintentionally triggering algorithm aversion or skepticism. Future experimental research could test more nuanced disclosure mechanisms, such as detailed process disclosures, percentage contributions of human versus AI effort, or interactive metadata standards. Investigating user reactions to varying disclosure formats will offer evidence-based guidelines to platforms and policymakers aiming to balance transparency with user engagement and authenticity perceptions.

Fifthly, the transferability of emotional and engagement patterns observed in this study should be tested across both alternative media formats and creative domains. While the current focus on Instagram art posts allows for high ecological validity in a visual, social-media-based setting, it does not guarantee that emotional responses generalize to other content categories, such as AI-generated writing, music, or performance-based art. Similarly, responses to AI content in entertainment, journalism, education, or political discourse may activate different emotional schemas, ethical appraisals, and trust dynamics. Future studies should thus examine how content genre and media modality interact with provenance disclosures to shape affective response.

Methodologically, enhancing the precision and inclusivity of affective classification remains crucial. The current dual-stage GPT–BERT classification pipeline is effective but limited by accuracy, linguistic biases, and emotional simplifications. Future studies should integrate human-in-the-loop adjudication or weak-supervision frameworks to improve classification accuracy, enable multi-label emotion annotations, and reduce linguistic and cultural biases inherited from existing corpora. Specifically, fine-tuning the BERT classifier on an Instagram-specific corpus could further enhance classification reliability. More refined emotional telemetry would substantially strengthen empirical estimates and theoretical interpretations related to algorithm aversion and emotional engagement.

Furthermore, the engagement depth classification introduced in this study via LLM provides valuable insight but requires further empirical testing and refinement. Future research should rigorously validate and improve the reliability and robustness of this engagement depth measure. This could involve qualitative validation with user interviews, experimental validation comparing classification results against user self-reports, or exploring alternative computational methods for reliably capturing engagement depth in user-generated content.

Additionally, demographic segmentation represents an unexplored but potentially influential moderator. While privacy constraints limited demographic insights in the present study, future research should

explicitly examine how user characteristics such as age, gender, cultural background, artistic expertise, or familiarity with technology, moderate emotional and engagement responses to AI-generated versus traditional content. Exploring these demographic nuances will enrich theoretical models and help platforms implement targeted transparency measures and engagement strategies.

In sum, pursuing these research directions will extend the current snapshot insights into dynamic, culturally nuanced, and methodologically robust understandings of user responses to algorithmic creativity. Integrating temporal, ecological, methodological, and demographic dimensions will significantly refine both theoretical frameworks and practical guidelines, shaping future policy, platform design, and scholarly inquiry.

7 Conclusion

Answering the primary research question, the results show that users respond to clearly labeled AI-generated art with diminished admiration, increased disapproval, and higher but shallower interaction compared to human-created art. The study addressed five central hypotheses, systematically testing the impact of content origin on discrete emotions, sentiment valence, behavioral engagement, and engagement depth, along with investigating whether users' prior exposure moderates these effects.

Empirically, the results demonstrate that clearly labeled AI-generated content elicits distinct emotional patterns characterized by diminished admiration and increased disapproval, confirming the hypothesis regarding discrete emotional responses (H1). Additionally, AI-generated content consistently receives more negative sentiment evaluations compared to traditionally created content, thus supporting the hypothesis on sentiment valence (H2). Despite these emotionally ambivalent or negative responses, AI-generated content attracts substantially higher superficial engagement in terms of likes, comments, and reshares, confirming the behavioral engagement hypothesis (H3). Moreover, the newly developed measure of engagement depth revealed that AI-generated art elicits significantly fewer deep and very deep comments, clearly distinguishing between superficial and meaningful user interactions and thus validating the hypothesis on engagement depth (H4). However, the moderation hypothesis (H5), examining the influence of users' prior exposure on their emotional responses, is not supported.

Theoretically, these findings contribute to ongoing discussions on algorithm aversion, signaling theory, and user engagement frameworks. Transparency disclosures act as affective signals that significantly reshape user perceptions, reframing the creative authenticity and legitimacy of digital artifacts. The dual emotional and engagement responses observed underscore a critical gap within existing engagement models, emphasizing the necessity of distinguishing between superficial and meaningful user interactions. Moreover, cultural dimensions were found to amplify these effects, with English-speaking users showing stronger negative emotional responses, suggesting that cultural factors and media discourse critically moderate user appraisals of AI-generated content.

Practically, the study’s outcomes inform platform policy and content strategy, emphasizing the nuanced effects of transparency disclosures. Rather than treating transparency merely as regulatory compliance, platforms and content creators must recognize it as a powerful cognitive-affective framing mechanism, strategically leveraging transparency to manage authenticity perceptions and emotional engagement. The observed disparity between high engagement volumes and shallow interaction depths also calls for revised metrics and engagement strategies, prioritizing qualitative indicators that capture emotional resonance and trust.

Methodologically, while robust in scale and ecological validity, the study acknowledged several critical limitations, notably regarding sampling biases from algorithmically curated content, the cross-sectional nature of data collection, the linguistic and cultural specificity of emotion labeling pipelines, and ethical constraints on demographic segmentation. Addressing these limitations through longitudinal studies, cross-platform validations, culturally sensitive classification methods, including fine-tuning classifiers on Instagram-specific corpora and demographic analyses represent important future research directions. Additionally, the engagement depth classification method introduced here should be rigorously tested and refined to establish its reliability and robustness as an analytical measure.

In personal reflection, the results highlight both the exciting potential and significant challenges associated with algorithmic creativity. AI-generated art can clearly capture attention and provoke curiosity yet simultaneously risks eroding deeper user engagement and perceived authenticity. The observed emotional ambivalence and superficial engagement present fundamental questions about the

evolving role of generative AI in creative domains. Beyond transparency, we must critically ask to what extent the benefits of AI-generated content justify potential compromises in user experience and emotional depth. As AI tools become increasingly embedded within cultural production, ensuring responsible, context-aware, and culturally sensitive transparency practices will be essential to fostering meaningful user experiences and maintaining trust in digital content ecosystems.

Ultimately, this thesis underscores the importance of critically evaluating how technological provenance shapes emotional and behavioral responses in digital spaces. By illuminating both the opportunities and pitfalls inherent to AI-generated content, it provides insights to guide future scholarship, policy, and platform practices toward more thoughtful integration and management of generative AI technologies.

References

Agresti, A. 2018. *Statistical Methods for the Social Sciences*. Pearson.

Alboqami, H. 2023. "Trust Me, I'm an Influencer! - Causal Recipes for Customer Trust in Artificial Intelligence Influencers in the Retail Industry," *Journal of Retailing and Consumer Services* (72).

Atkinson, D. P., and Barker, D. R. 2023. "Ai and the Social Construction of Creativity," *Convergence* (29:4), pp. 1054-1069.

Bach, T. A., Amna, K., Harry, H., Gabriela, B., and and Sousa, S. 2022. "A Systematic Literature Review of User Trust in Ai-Enabled Systems: An Hci Perspective," *International Journal of Human–Computer Interaction* (40:5), pp. 1251-1266.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. 2012. "The Role of Social Networks in Information Diffusion," *'12 - Proceedings of the 21st Annual Conference on World Wide Web*).

Banh, L., and Strobel, G. 2023. "Generative Artificial Intelligence," *Electronic Markets* (33:1).

Barrett, L. F. 2006. "Solving the Emotion Paradox: Categorization and the Experience of Emotion," *Personality and Social Psychology Review* (10:1), pp. 20-46.

Bauer, K., Jussupow, E., Heigl, R., Vogt, B., and Hinz, O. 2024. "All Just in Your Head? Unraveling the Side Effects of Generative Ai Disclosure in Creative Task," *SSRN Electronic Journal*).

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., Christensen, A., and Seli, P. 2023. "Humans Versus Ai: Whether and Why We Prefer Human-Created Compared to Ai-Created Artwork," *Cognitive Research: Principles and Implications* (8:1), p. 42.

Berlyne, D. E. 1960. *Conflict, Arousal, and Curiosity*. New York, NY, US: McGraw-Hill Book Company.

Boden, M. A. 1998. "Creativity and Artificial Intelligence," *Artificial Intelligence* (103:1), pp. 347-356.

Boitel, E., Mohasseb, A., and Haig, E. 2024. "A Comparative Analysis of Gpt-3 and Bert Models for Text-Based Emotion Recognition: Performance, Efficiency, and Robustness." pp. 567-579.

Bota, P. J., Wang, C., Fred, A. L. N., and Silva, H. P. D. 2019. "A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals," *IEEE Access* (7), pp. 140990-141020.

Brodie, R. J., Hollebeek, L., Juric, B., and Ilic, A. 2011. "Customer Engagement: Conceptual Domain, Fundamental Propositions, and Implications for Research," *Journal of Service Research* (17), pp. 1-20.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. 2020. "Language Models Are Few-Shot Learners," *Advances in neural information processing systems* (33), pp. 1877-1901.

C2PA. 2024. "Content Credentials : C2pa Technical Specification." Retrieved 18.04., 2025, from https://c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html

Cameron, A. C., and Trivedi, P. K. 2013. *Regression Analysis of Count Data*, (2 ed.). Cambridge: Cambridge University Press.

Cetinic, E., and She, J. 2021. "Understanding and Creating Art with Ai: Review and Outlook," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (18), pp. 1 - 22.

Cheng, K.-T., Chang, K., and Tai, H.-W. 2022. "Ai Boosts Performance but Affects Employee Emotions," *Information Resources Management Journal* (35), pp. 1-18.

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., and de Oliveira, N. 2023. "Worldwide Ai Ethics: A Review of 200 Guidelines and Recommendations for Ai Governance," *Patterns* (4:10), p. 100857.

Cowen, A. S., and Keltner, D. 2017. "Self-Report Captures 27 Distinct Categories of Emotion Bridged by Continuous Gradients," *Proceedings of the National Academy of Sciences* (114:38), pp. E7900-E7909.

Creswell, J. W. 2014. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications.

D'Andrea, A., Ferri, F., Grifoni, P., and Guzzo, T. 2015. "Approaches, Tools and Applications for Sentiment Analysis Implementation," *International Journal of Computer Applications* (125), pp. 26-33.

Daft, R., and Lengel, R. 1986. "Organizational Information Requirements, Media Richness and Structural Design," *Management Science* (32), pp. 554-571.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. 2020. *Goemotions: A Dataset of Fine-Grained Emotions*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171-4186.

Dewaele, J.-M. 2010. *Emotions in Multiple Languages*. Springer.

Di Gangi, P., and Wasko, M. 2016. "Social Media Engagement Theory," *Journal of Organizational and End User Computing* (28), pp. 53-73.

Dietvorst, B., Simmons, J., and Massey, C. 2014. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of experimental psychology. General* (144).

Doshi, A. R., and Hauser, O. P. 2023. "Generative Ai Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content," *Science Advances* (10:28).

Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., and Welsch, R. 2023. *The Ai Ghostwriter Effect: Users Do Not Perceive Ownership of Ai-Generated Text but Self-Declare as Authors*.

Eerola, T., and Vuoskoski, J. K. 2011. "A Comparison of the Discrete and Dimensional Models of Emotion in Music," *Psychology of Music* (39:1), pp. 18-49.

Ekman, P. 1992. "An Argument for Basic Emotions," *Cognition and Emotion* (6:3-4), pp. 169-200.

European Parliament, and Madiega, T. 2024. "Artificial Intelligence Act," in: *Briefing*.

European Parliament and Council. 2024. "Regulation (Eu) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (Ec) No 300/2008, (Eu) No 167/2013, (Eu) No 168/2013, (Eu) 2018/858, (Eu) 2018/1139 and (Eu) 2019/2144 and Directives 2014/90/Eu, (Eu) 2016/797 and (Eu) 2020/1828 (Artificial Intelligence Act)." Brussels: Official Journal of the European Union.

European Union. 2016. "Regulation (Eu) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/Ec (General Data Protection Regulation) ", T.E.P.A.T.C.O.T.E. UNION (ed.).

Field, A., Miles, J., and Field, Z. 2012. *Discovering Statistics Using R*. SAGE Publications.

Gabbiadini, A., Ognibene, D., Baldissarri, C., and Manfredi, A. 2024. "The Emotional Impact of Generative Ai: Negative Emotions and Perception of Threat," *Behaviour & Information Technology*), pp. 1-18.

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., and Spaulding, R. 2021. "Attachment and Trust in Artificial Intelligence," *Computers in Human Behavior* (115).

Gkinko, L., and Elbanna, A. 2022. "Hope, Tolerance and Empathy: Employees' Emotions When Using an Ai-Enabled Chatbot in a Digitalised Workplace," *Information Technology & People*).

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*. MIT Press.

Google Research. 2021. "Goemotions: A Dataset for Fine-Grained Emotion Classification," D.A.a.J. Ko (ed.). https://research.google/.

Grimmer, J., and Stewart, B. M. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* (21:3), pp. 267-297.

Guzik, E. E., Byrge, C., and Gilde, C. 2023. "The Originality of Machines: Ai Takes the Torrance Test," *Journal of Creativity* (33:3).

Harmon-Jones, C., Bastian, B., and Harmon-Jones, E. 2016. "The Discrete Emotions Questionnaire: A New Tool for Measuring State Self-Reported Emotions," *PLOS ONE* (11:8), p. e0159915.

Hilbe, J. M. 2011. *Negative Binomial Regression*. Cambridge University Press.

Hofstede, G. 2011. "Dimensionalizing Cultures: The Hofstede Model in Context," *Online readings in psychology and culture* (2:1), p. 8.

Hollebeek, L. D., Glynn, M. S., and Brodie, R. J. 2014. "Consumer Brand Engagement in Social Media: Conceptualization, Scale Development and Validation," *Journal of Interactive Marketing* (28:2), pp. 149-165.

Instagram. 2025. "Instagram Search." Retrieved 06.01., 2025, from https://www.instagram.com/

Jahan, R. I., Fan, H., Chen, H., and Feng, Y. 2024. *Unlocking Cross-Lingual Sentiment Analysis through Emoji Interpretation: A Multimodal Generative Ai Approach*.

Janiesch, C., Zschech, P., and Heinrich, K. 2021. "Machine Learning and Deep Learning," *Electronic Markets* (31:3), pp. 685-695.

Jiang, Y., Li, X., Luo, H., Yin, S., and Kaynak, O. 2022. "Quo Vadis Artificial Intelligence?," *Discover Artificial Intelligence* (2:1), p. 4.

Jünger, J. 2023. *Scraping Social Media Data as Platform Research: A Data Hermeneutical Perspective*. DEU

Berlin:

Katz, E., Blumler, J. G., and Gurevitch, M. 1973. "Uses and Gratifications Research," *Public Opinion Quarterly* (37:4), pp. 509-523.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., and Kazienko, P. 2023. "Chatgpt: Jack of All Trades, Master of None," *Information Fusion* (99), p. 101861.

Larsson, S., and Heintz, F. 2020. "Transparency in Artificial Intelligence," *Internet policy review* (9:2), pp. 1-16.

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., and Wagner, C. 2020. "Computational Social Science: Obstacles and Opportunities," *Science* (369:6507), pp. 1060-1062.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. 2019. *Roberta: A Robustly Optimized Bert Pretraining Approach*.

Logg, J. M., Minson, J. A., and Moore, D. A. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes* (151), pp. 90-103.

Lowe, S. 2023. "Evaluating the Roberta-Base-Go_Emotions Model." Retrieved 08.05, 2025, from https://github.com/samlowe/go_emotions-dataset/blob/main/eval-roberta-base-go_emotions.ipynb

Lowe, S. 2024. "Roberta-Base-Go_Emotions (Revision 58b6c5b)." Retrieved 08.05, 2025, from https://huggingface.co/SamLowe/roberta-base-go_emotions

Lukyanenko, R., Maass, W., and Storey, V. C. 2022. "Trust in Artificial Intelligence: From a Foundational Trust Framework to Emerging Research Opportunities," *Electronic Markets* (32:4), pp. 1993-2020.

Lund, B., Orhan, Z., Mannuru, N. R., Bevara, R. V. K., Porter, B., Vinaih, M. K., and Bhaskara, P. 2025. "Standards, Frameworks, and Legislation for Artificial Intelligence (Ai) Transparency," *AI and Ethics*).

Mauss, I. B., and Robinson, M. D. 2009. "Measures of Emotion: A Review," *Cognition and Emotion* (23:2), pp. 209-237.

Mazaheri, E., Lagzian, M., and Hemmat, Z. 2020. "Research Directions in Information Systems Field, Current Status and Future Trends: A Literature Analysis of Ais Basket of Top Journals," *Australasian Journal of Information Systems* (24:0).

McHugh, M. 2012. "Interrater Reliability: The Kappa Statistic," *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB* (22), pp. 276-282.

Meta. 2024. "Our Approach to Labeling Ai-Generated Content and Manipulated Media." Retrieved 20.12.2024, 2024, from https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media

Meta Help Center. 2025. "How to Identify Ai Content on Meta Products." Retrieved 18.04., 2025, from https://www.meta.com/help/artificial-intelligence/1783222608822690/?srsltid=AfmBOoozl1S9dBBck9a7FmhMZ6LZ0XyUhIDI5vThw9Oyhj1at7YhWsWM

Mori, M., MacDorman, K. F., and Kageki, N. 2012. "The Uncanny Valley [from the Field]," *IEEE Robotics & Automation Magazine* (19:2), pp. 98-100.

Nandwani, P., and Verma, R. 2021. "A Review on Sentiment Analysis and Emotion Detection from Text," *Social Network Analysis and Mining* (11).

Nissenbaum, H. 2010. "Privacy in Context: Technology, Policy, and the Integrity of Social Life," *Bibliovault OAI Repository, the University of Chicago Press*).

Niu, M., Jaiswal, M., and Provost, E. M. 2024. "From Text to Emotion: Unveiling the Emotion Annotation Capabilities of Llms," *ArXiv* (abs/2408.17026).

O'Day, E. B., and Heimberg, R. G. 2021. "Social Media Use, Social Anxiety, and Loneliness: A Systematic Review," *Computers in Human Behavior Reports* (3), p. 100070.

Ooi, K.-B., Wei-Han, T. G., Mostafa, A.-E., A., A.-S. M., Alexandru, C., Amrita, C., K., D. Y., Tzu-Ling, H., Kumar, K. A., Voon-Hsien, L., Xiu-Ming, L., Adrian, M., Patrick, M., Emmanuel, M., Neeraj, P., Ramakrishnan, R., P., R. N., Prianka, S., Anshuman, S., I., T. C.-., Fosso, W. S., and and Wong, L.-W. 2025. "The Potential of Generative Artificial Intelligence across Disciplines: Perspectives and Future Directions," *Journal of Computer Information Systems* (65:1), pp. 76-107.

Park, J., Oh, C., and Kim, H. Y. 2024. "Ai Vs. Human-Generated Content and Accounts on Instagram: User Preferences, Evaluations, and Ethical Considerations," *Technology in Society* (79).

Peters, Y., Nehls, P., and Thimm, C. 2023. "Plattformforschung Mit Instagram-Daten – Eine Übersicht Über Analytische Zugänge, Digitale Erhebungsverfahren Und Forschungsethische Perspektiven in Zeiten Der Apicalypse," *Publizistik* (68:2), pp. 225-239.

Pittman, M., and Reich, B. 2016. "Social Media and Loneliness: Why an Instagram Picture May Be Worth More Than a Thousand Twitter Words," *Computers in Human Behavior* (62), pp. 155-167.

Plutchik, R. 1980. "Chapter 1 - a General Psychoevolutionary Theory of Emotion," in *Theories of Emotion,* R. Plutchik and H. Kellerman (eds.). Academic Press, pp. 3-33.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. 2018. "Improving Language Understanding by Generative Pre-Training,").

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. 2020. "Snorkel: Rapid Training Data Creation with Weak Supervision," *The VLDB Journal* (29:2), pp. 709-730.

Recker, J. 2021. *Scientific Research in Information Systems: A Beginner's Guide. Second Edition*.

Roose, K. 2022. "An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy." Retrieved 06.01.2025, 2025, from https://www.nytimes.com/2022/09/02/technology/ai-artificialintelligence-artists.html.

Russell, J. 1980. "A Circumplex Model of Affect," *Journal of Personality and Social Psychology* (39), pp. 1161-1178.

Saltarella, M., Desolda, G., and Lanzilotti, R. 2021. "Privacy Design Strategies and the Gdpr: A Systematic Literature Review." pp. 241-257.

Schivinski, B., Christodoulides, G., and Dabrowski, D. 2016. "Measuring Consumers' Engagement with Brand-Related Social-Media Content: Development and Validation of a Scale That Identifies Levels of Social-Media Engagement with Brands," *Journal of Advertising Research* (56).

Sengar, S. S., Hasan, A. B., Kumar, S., and Carroll, F. 2024. "Generative Artificial Intelligence: A Systematic Review and Applications," *Multimedia Tools and Applications*).

Smith, C. A., and Lazarus, R. S. 1993. "Appraisal Components, Core Relational Themes, and the Emotions," *Cognition and Emotion* (7:3-4), pp. 233-269.

Spence, M. 1973. "Job Market Signaling," *The Quarterly Journal of Economics* (87:3), pp. 355-374.

The White House. 2023. "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." Washington, D.C.: Federal Register / Government Publishing Office, pp. 75191–75226.

Trunfio, M., and Rossi, S. 2021. "Conceptualising and Measuring Social Media Engagement: A Systematic Literature Review," *Italian Journal of Marketing* (2021:3), pp. 267-292.

Vasist, P., and Krishnan, S. 2022. "Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research," *Communications of the Association for Information Systems* (51), pp. 590-636.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. "Attention Is All You Need," in: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California, USA: Curran Associates Inc., pp. 6000–6010.

Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., and Kross, E. 2017. "Do Social Network Sites Enhance or Undermine Subjective Well-Being? A Critical Review: Do Social Network Sites Enhance or Undermine Subjective Well-Being?," *Social Issues and Policy Review* (11), pp. 274-302.

Wickens, T. D. 2014. *Multiway Contingency Tables Analysis for the Social Sciences*. Taylor & Francis.

Wilcox, R. 2012. *Introduction to Robust Estimation and Hypothesis Testing. 3rd Ed*.

Wu, Z., Ji, D., Yu, K., Zeng, X., Wu, D., and Shidujaman, M. 2021. "Ai Creativity and the Human-Ai Co-Creation Model," *Human-Computer Interaction. Theory, Methods and Tools,* M. Kurosu (ed.), Cham: Springer International Publishing, pp. 171-190.

Zhou, E., and Lee, D. 2024. "Generative Artificial Intelligence, Human Creativity, and Art," *PNAS Nexus* (3:3).

Appendix A: Declaration on the use of GenAI tools

In the preparation of this paper, I have used following tools based on generative artificial intelligence (GenAI):

1. ChatGPT
2. BERT

I further declare that

- I have labeled the content taken from the GenAI tools listed above with my details in the table below,
- I have verified that the content generated by the above-mentioned GenAI tools and adapted by me is factually correct,
- I am aware that, as the author of this work, I am responsible for the information and the statements made in it, and
- I am aware that violating the disclosure of the use of generative AI in my work is a deception and leads to an evaluation with an insufficient grade.

I have used the above-mentioned AI systems as indicated below.

Areas of contribution	**AI tool(s) used**	**Description of the manner of use and compliance with good scientific practice (if applicable, please indicate the section of the thesis)**
Development and conception of the research project	1	Initial review of various research approaches
Identification of literature		
Synthesizing of literature	1	Analysis, summary, consolidation, review of identified literature
Structuring the text	1	Initial bainstorming, review, revision of structure
Formulation of text	1	Re-formulation of text and passages based on context and inputs. Stylistic rework and revision
Revision of text	1	General revision, stylistic revision, readability, grammar and language
Creation of visualizations		
Further contributions	1, 2	Coding, Code review, Data overview, Data Analysis, Data Classification (see Pipeline)

Appendix B: Final LLM instruction prompt

You are an expert in analyzing social media comments. Your task is to analyze a single Instagram comment and classify it into emotional and engagement-related categories.

Please return your answer in this JSON format:

```
{
  "sentiment": ...,
  "emotion_go": ...,
  "emotion_go_confidence": ...,
  "engagement_depth": ...,
  "sarcasm": ...,
  "language": ...
}
```

Important instructions:

- Always return exactly ONE value for "emotion_go". Do NOT return a list or multiple values.
- If multiple emotions seem relevant, choose the single most dominant one.
- All output must follow strict JSON formatting and match the expected data types.

Definitions:

- sentiment: ["Positive", "Neutral", "Negative"]
- emotion_go: [admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral]
- emotion_go_confidence: A float between 0 and 1
- engagement_depth:
    - Superficial: short, emoji-only, or generic
    - Moderate: brief, specific compliment
    - Deep: thoughtful, critique, or question
    - Very Deep: personal story, emotional reflection
- sarcasm: ["Yes", "No", "Maybe"]
- language: ISO language code, or "Emoji-Only", "Emoji-Dominant"

Here are a few examples:

Example1:

Comment: "🔥🔥🔥🔥🔥"

{

"engagement_depth": "Superficial"

}

Example2:

Comment: "It looks great! I love the cools in the lightning"

{

"engagement_depth": "Moderate"

}

Example3:

Comment: "What white pen do you use. I've been looking for a good one for a long time"

{

"engagement_depth": "Deep"

}

Example4:

Comment: "You are a huge inspiration. I am from right outside […] but now live in […]. I'm tight with the guys at […] and they said you came in a while back. I was in high school when I purchased your paperback (with your cover). Yourself and many others have influenced me since getting back into painting about six years ago. […] So funny how these things work out. Take care!"

{

"engagement_depth": "Very Deep"

}

"""

Appendix C: Requirements

```
numpy==1.26.4

pandas==2.2.1

scipy==1.15.3

scikit-learn==1.6.1

statsmodels==0.14.4

matplotlib==3.8.4

seaborn==0.13.2

tqdm==4.67.1

torch==2.7.0

transformers==4.51.3

huggingface-hub==0.31.2

openai==1.68.2

tabulate==0.9.0
```

Appendix D: Summarized distributions of classifier confidence

Average BERT confidence across English subset: 0.808

Confidence distribution per BERT-label:

emotion_bert	count	mean	std	min	25%	50%	75%	max
admiration	4528	0.88	0.121	0.152	0.859	0.931	0.952	0.973
neutral	4371	0.823	0.195	0.127	0.711	0.943	0.963	0.973
love	2200	0.877	0.115	0.201	0.841	0.923	0.949	0.972
curiosity	710	0.622	0.115	0.226	0.545	0.61	0.692	0.917
excitement	314	0.523	0.15	0.13	0.426	0.52	0.635	0.825
approval	314	0.553	0.159	0.128	0.458	0.534	0.659	0.926
surprise	273	0.648	0.158	0.248	0.526	0.709	0.778	0.895
gratitude	223	0.897	0.177	0.229	0.912	0.974	0.987	0.994
desire	215	0.654	0.155	0.173	0.561	0.696	0.774	0.875
sadness	141	0.682	0.202	0.212	0.518	0.718	0.864	0.93
amusement	133	0.715	0.214	0.124	0.575	0.792	0.886	0.958
joy	124	0.701	0.165	0.246	0.6	0.74	0.844	0.924
disapproval	105	0.615	0.146	0.263	0.506	0.615	0.735	0.885
annoyance	104	0.465	0.123	0.227	0.37	0.458	0.551	0.719
anger	89	0.627	0.149	0.291	0.502	0.648	0.763	0.839
confusion	86	0.643	0.155	0.366	0.521	0.622	0.753	0.931
caring	67	0.687	0.162	0.256	0.613	0.719	0.803	0.903
disappointment	59	0.491	0.145	0.244	0.37	0.496	0.588	0.818
optimism	59	0.661	0.24	0.092	0.49	0.746	0.859	0.952
disgust	33	0.46	0.185	0.171	0.328	0.4	0.595	0.834
fear	31	0.724	0.197	0.251	0.599	0.801	0.864	0.923
realization	26	0.541	0.139	0.361	0.422	0.536	0.594	0.864
remorse	21	0.692	0.116	0.418	0.646	0.739	0.775	0.823
embarrassment	2	0.372	0.004	0.368	0.37	0.372	0.373	0.375
nervousness	1	0.643	nan	0.643	0.643	0.643	0.643	0.643
pride	1	0.462	nan	0.462	0.462	0.462	0.462	0.462

Average GPT confidence across English subset: 0.848

Confidence distribution per GPT-label:

emotion_go	count	mean	std	min	25%	50%	75%	max
admiration	6091	0.862	0.025	0.7	0.85	0.85	0.9	0.95
curiosity	1568	0.809	0.058	0.6	0.8	0.85	0.85	0.85
approval	1291	0.864	0.035	0.7	0.85	0.85	0.9	0.9
neutral	1044	0.835	0.048	0.8	0.8	0.8	0.9	1
confusion	860	0.774	0.061	0.6	0.7	0.8	0.8	0.85
amusement	567	0.843	0.032	0.7	0.85	0.85	0.85	0.9
joy	559	0.864	0.026	0.8	0.85	0.85	0.9	0.9
disapproval	366	0.848	0.022	0.7	0.85	0.85	0.85	0.9
love	306	0.899	0.024	0.8	0.9	0.9	0.9	0.95
excitement	296	0.874	0.027	0.8	0.85	0.85	0.9	0.95
desire	237	0.851	0.012	0.8	0.85	0.85	0.85	0.9
caring	189	0.845	0.034	0.7	0.85	0.85	0.85	0.95
sadness	182	0.852	0.021	0.7	0.85	0.85	0.85	0.9
anger	159	0.849	0.024	0.75	0.85	0.85	0.85	0.9
surprise	126	0.842	0.039	0.7	0.8	0.85	0.85	0.9
gratitude	126	0.878	0.027	0.8	0.85	0.9	0.9	0.95
disappointment	100	0.836	0.031	0.7	0.85	0.85	0.85	0.85
annoyance	47	0.821	0.046	0.7	0.8	0.85	0.85	0.85
fear	43	0.833	0.031	0.75	0.8	0.85	0.85	0.9
disgust	18	0.842	0.026	0.8	0.85	0.85	0.85	0.9
grief	15	0.857	0.018	0.85	0.85	0.85	0.85	0.9
relief	12	0.85	0	0.85	0.85	0.85	0.85	0.85
optimism	11	0.85	0	0.85	0.85	0.85	0.85	0.85
realization	8	0.844	0.018	0.8	0.85	0.85	0.85	0.85
pride	7	0.864	0.024	0.85	0.85	0.85	0.875	0.9

Appendix E: Full Emotion Effects

	OR	*OR_CI_low*	*OR_CI_high*	*p_value*	*Δ(APP)*	*Δ_CI_low*	*Δ_CI_high*
admiration	0.90	0.85	0.95	0.001	-0.049	-0.0498	-0.0498
amusement	1.94	1.72	2.20	0.0	0.0173	0.0174	0.0174
anger	1.38	1.01	1.89	0.038	0.001	0.001	0.001
annoyance	1.59	1.11	2.29	0.012	0.0010	0.0011	0.0011
approval	1.00	0.92	1.08	0.991	-0.000	-0.0004	-0.0004
caring	1.70	1.48	1.94	0.0	0.0101	0.0102	0.0102
confusion	1.28	1.10	1.49	0.001	0.0033	0.0033	0.0033
curiosity	0.62	0.55	0.70	0.0	-0.009	-0.01	-0.01
desire	0.67	0.54	0.83	0.0	-0.002	-0.0024	-0.0024
disappointment	0.75	0.50	1.12	0.163	-0.000	-0.0005	-0.0005
disapproval	3.22	2.31	4.49	0.0	0.0043	0.0043	0.0043
disgust	1.11	0.70	1.76	0.636	0.0001	0.0001	0.0001
excitement	0.51	0.44	0.59	0.0	-0.009	-0.0096	-0.0096
fear	0.97	0.60	1.58	0.934	-9.999	-0.0	-0.0
gratitude	1.09	0.92	1.31	0.299	0.0009	0.0008	0.0008
grief	3.36	1.32	8.55	0.011	0.0006	0.0006	0.0006
joy	1.74	1.55	1.94	0.0	0.017	0.017	0.017
love	1.12	1.05	1.19	0.0	0.0181	0.0182	0.0182
optimism	0.49	0.31	0.78	0.003	-0.000	-0.0009	-0.0009
pride	0.09	0.01	0.52	0.006	-0.000	-0.0004	-0.0004
realization	0.46	0.23	0.94	0.035	-0.000	-0.0004	-0.0004
remorse	0.67	0.28	1.60	0.371	-0.000	-0.0001	-0.0001
sadness	1.34	1.13	1.59	0.001	0.0031	0.0032	0.0032
surprise	0.82	0.69	0.98	0.032	-0.001	-0.0019	-0.0019

Appendix F: English-Subset Emotion Effects

	OR	*OR_CI_low*	*OR_CI_high*	*p_value*	*Δ(APP)*	*Δ_CI_low*	*Δ_CI_high*
disapproval	2.37	1.46	3.83	0.0	0.009	0.009	0.009
annoyance	1.61	1.04	2.49	0.032	0.0069	0.0069	0.0069
fear	1.23	0.57	2.63	0.584	0.00149	0.0016	0.0016
anger	1.17	0.75	1.82	0.478	0.0042	0.0042	0.0042
caring	0.88	0.54	1.45	0.638	0.0019	0.002	0.002
amusement	0.81	0.57	1.15	0.246	0.0031	0.0031	0.0031
confusion	0.77	0.50	1.19	0.249	0.0017	0.0017	0.0017
joy	0.76	0.54	1.09	0.141	0.0025	0.0025	0.0025
remorse	0.76	0.32	1.82	0.551	0.00039	0.0004	0.0004
sadness	0.75	0.53	1.05	0.096	0.00249	0.0025	0.0025
approval	0.73	0.58	0.92	0.01	0.0052	0.0052	0.0052
disappointment	0.62	0.37	1.03	0.066	0.00029	0.0003	0.0003
curiosity	0.59	0.50	0.69	0.0	0.00089	0.0009	0.0009
realization	0.58	0.27	1.23	0.16	0.0	0.0	0.0
desire	0.53	0.40	0.70	0.0	-0.0014	-0.0014	-0.0014
excitement	0.50	0.40	0.63	0.0	-0.0030	-0.0031	-0.0031
gratitude	0.50	0.38	0.66	0.0	-0.0021	-0.0022	-0.0022
optimism	0.49	0.30	0.82	0.006	-0.0006	-0.0007	-0.0007
disgust	0.48	0.24	0.96	0.04	-0.0004	-0.0004	-0.0004
surprise	0.48	0.37	0.61	0.0	-0.0036	-0.0036	-0.0036
love	0.40	0.36	0.45	0.0	-0.0554	-0.0554	-0.0554
admiration	0.38	0.34	0.41	0.0	-0.1341	-0.1342	-0.1342

Appendix G: Pairwise chi-square tests between engagement depth levels

Contrast	χ^2	p (Holm-corrected)
Superficial vs Moderate	1906.03	< .001
Superficial vs Deep	377.22	< .001
Superficial vs Very Deep	28.45	< .001
Moderate vs Deep	4.37	.110
Moderate vs Very Deep	0.06	.810
Deep vs Very Deep	0.87	.701