TU WIEN Informatics

# Kontext statt Kategorien - Implementierung der Theorie of Constructed Emotion

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieur**

im Rahmen des Studiums

**Wirtschaftsinformatik**

eingereicht von


**Nils Klüwer, M.Sc, B.Sc**
Matrikelnummer 12229263

an der Fakultät für Informatik

der Technischen Universität Wien


Betreuung: Assistant Prof. Mag.a rer.nat. Dr.in techn. Julia Neidhardt
Mitwirkung: Projektass.in Mag.a rer.nat. Dr.in rer.nat. Irina Nalis-Neuner

Wien, 31. März 2025

| Nils Klüwer | Julia Neidhardt |
|---|---|

Informatics

# Context over Categories - Implementing the Theory of Constructed Emotion

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieur**

in

**Business Informatics**

by

**Nils Klüwer, M.Sc, B.Sc**
Registration Number 12229263

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Mag.a rer.nat. Dr.in techn. Julia Neidhardt
Assistance: Projektass.in Mag.a rer.nat. Dr.in rer.nat. Irina Nalis-Neuner

Vienna, March 31, 2025

Nils Klüwer    Julia Neidhardt

# Erklärung zur Verfassung der Arbeit

Nils Klüwer, M.Sc, B.Sc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 31. März 2025

Nils Klüwer

# Danksagung

Zuerst möchte ich mich ganz herzlich beim Christian Doppler Labor für Recommender Systems bedanken. Für die Betreuung meiner Arbeit und die wertvolle Unterstützung möchte ich mich bei Irina und Julia herzlich bedanken. Ihr konstruktives Feedback war stets zielführend und hat wesentlich zur Qualität dieser Arbeit beigetragen. Besonders hervorheben möchte ich ihr Engagement für den interdisziplinären Ansatz meines Projekts. Ihre Offenheit für innovative Ideen und die Förderung dieses Ansatzes haben es ermöglicht, technisch anspruchsvolle Lösungen zu entwickeln. Darüber hinaus danke ich Ihnen für die Ermutigung, meine Ergebnisse auch einem größeren Fachpublikum zu präsentieren, insbesondere durch die Einreichung als Late-Breaking Work für die CHI 2025. Ihre Motivation, den Wert meiner Forschung über den Rahmen der Diplomarbeit hinaus zu erkennen und zu unterstützen, habe ich sehr geschätzt.

Mein Dank gilt auch Der Standard, der österreichischen Tageszeitung, für die Bereitstellung der Daten. Ohne diese wertvolle Datengrundlage wäre diese Arbeit in dieser Form nicht realisierbar gewesen.

Ebenso möchte ich mich bei den fünf Interviewteilnehmern Inga, Ben, Sophia, Paul und Chris für ihre Teilnahme an den Interviews bedanken. Ihre Bereitschaft, sich Zeit zu nehmen und ihre Perspektiven zu teilen, war sehr wertvoll für diese Arbeit.

Ein besonderer Dank geht an die Abteilung Research & Innovation der REWE digital GmbH für die Bereitstellung einer sicheren Entwicklungsumgebung mit Zugangs zu den Sprachmodellen und für Übernahme der Kosten die durch die Sprachmodell-Ausführung entstanden sind. Diese Unterstützung hat die praktische Umsetzung und den Umfang meiner Arbeit in entscheidendem Maße ermöglicht.

All diese Unterstützung hat maßgeblich zum Gelingen dieser Arbeit beigetragen. Vielen Dank!

# Acknowledgements

First and foremost, I would like to express my sincere gratitude to the Christian Doppler Laboratory for Recommender Systems. For their supervision of my work and their invaluable support, I would like to sincerely thank Irina and Julia. Their constructive feedback was consistently insightful and significantly enhanced the quality of this Thesis. I particularly want to emphasize their commitment to the interdisciplinary approach of this project. Their openness to innovative ideas and their encouragement of this approach enabled the development of technically sophisticated solutions. Furthermore, I thank them for their encouragement to present my findings to a wider academic audience, especially through the submission of a Late-Breaking Work to CHI 2025. I greatly appreciate their motivation to recognize and support the value of my research beyond the scope of this Thesis.

My gratitude also goes to Der Standard, the Austrian daily newspaper, for sharing their data with us. Without this valuable data basis, this work would not have been feasible in this form.

I would also like to thank the five interview participants, Inga, Ben, Sophia, Paul, and Chris, for their participation in the interviews. Their willingness to dedicate their time and share their perspectives was invaluable to this work.

Special thanks go to the Research & Innovation department of REWE digital GmbH for providing a secure environment to access the LLMs and for covering the costs of LLM execution. This support significantly enabled the practical implementation and scope of my work.

All of this support was instrumental in the success of this Thesis. Thank you very much again!

# Kurzfassung

Die vorliegende Arbeit untersucht die Emotionsanalyse, ein bedeutendes Forschungsfeld mit Anwendungen, die von der Moderation von Online-Inhalten bis hin zu personalisierten Systemen reichen. Traditionelle Modelle, wie Ekmans Theorie universeller Emotionen, fassen Emotionen oft in statische, vordefinierte Kategorien, deren komplexe und kontextabhängige Natur vernachlässigt wird. Diese Arbeit stützt sich hingegen auf Lisa Feldman Barretts Theorie der konstruierten Emotionen, um einen neuartigen, kontextsensiblen Analyseansatz zu entwickeln. Ein zentraler Beitrag ist die Einführung der "context sphere" – eines personalisierten, datenbasierten Konstrukts, das aus Nutzerverhaltensdaten abgeleitet wird. Dieses ermöglicht eine dynamische Modellierung von Emotionen als emergente, kontextabhängige Phänomene. Im Rahmen der Design Science Research-Methodik wird eine Emotionsanalyse-Pipeline entwickelt, die fortschrittliche Techniken des LLM-Prompting, wie beispielsweise Role-Play Prompting, Controlled Generation und Meta-Prompting, integriert. Eine Fallstudie zur Online-Content-Moderation sowie eine Evaluation durch Leitfaden geführte Interviews und einen LLM-as-a-Judge-Ansatz zeigen, dass die entwickelte Pipeline eine nuancierte und kontextbewusste Analyse von Emotionen ermöglicht, die über die Begrenzungen traditioneller Modelle hinausgeht, ohne sich ausschließlich auf fest definierte Emotionslabels zu stützen. Besonders bemerkenswert ist, dass alle fünf Interviewteilnehmer eine vollständige Übereinstimmung ihrer Interpretationen mit der durch die Pipeline generierten Analyse zeigten, was darauf hindeutet, dass eine kontextbezogene Emotionsanalyse möglich ist. Diese Forschung schlägt eine Brücke zwischen den Fortschritten in den Kognitionswissenschaften und der Informatik, eröffnet neue Wege für menschenzentrierte, ethische und effektive Anwendungen und weist auf zukünftige Forschungsrichtungen zur Weiterentwicklung der "context sphere" und der LLM-basierten Analyseverfahren hin.

# Abstract

Emotion analysis is a critical research area with applications ranging from content moderation to personalized systems. Despite its importance, many approaches rely on traditional models, such as Ekman's universal emotions theory, which reduces emotions to static, predefined categories. This oversimplification neglects the complexity and contextual variability of human emotions. Drawing on Lisa Feldman Barrett's Theory of Constructed Emotion introduces a novel, context-aware approach to emotion analysis. A central contribution is the development of the "context sphere", a personalized construct derived from user behavior data. To our knowledge, this is the first operationalization of Barrett's theory for computational methods, enabling the dynamic modeling of emotions as emergent and context-dependent phenomena. The Design Science Research methodology is employed to develop a context-aware emotion analysis pipeline, which utilizes advanced Large Language Model (LLM) prompting strategies, including Role-Play prompting and Controlled Generation, to align with constructed emotion principles. A case study in online content moderation demonstrates the feasibility of this approach, showing that the "context sphere" facilitates nuanced and contextually aware emotion analyses that surpass traditional model limitations. Furthermore, a Semi-Structured Interview evaluation with five participants revealed strong alignment, with all participants agreeing that the LLM-generated analysis closely matched their perceptions based on the "context sphere". Although minor disagreements regarding individual words were noted, there was no consistency across participants in these discrepancies. These findings suggest that the pipeline successfully analyzes user emotions using the Theory of Constructed Emotion without relying on distinct labels. This work bridges advancements in cognitive science and computer science to propose a novel, context-aware approach to emotion analysis. Challenges in evaluating such analyses without a definitive ground truth are discussed, and future directions include refining the "context sphere," enhancing LLM-guided methodologies for emotion analysis, and integrating different user studies. This research lays the groundwork for developing more nuanced, context-sensitive emotion analysis systems, opening new avenues for human-centered, ethical, and effective applications in technology.

# Contents

# CHAPTER 1

# Introduction

## 1.1 Preface and Motivation

Emotion analysis has become an increasingly significant area of research, with applications spanning sentiment analysis, content moderation, and human-centered adaptation. Despite its prevalence, much of the existing work in this domain continues to rely on traditional models, such as Paul Ekman's theory of universal emotions [EF71, AM20]. This framework, while widely adopted, simplifies the complexity of human emotion into a set of discrete categories. This simplification has also shaped recent work in informatics, where emotion analysis often relies on predefined categories, particularly in natural language processing (NLP) applications such as social media analysis. These approaches frequently focus on individual words, sentences, and basic sentiments [BCCNEA20, YSZ17]. Moreover, resources like the NRC Word-Emotion Association Lexicon exemplify this trend by relying on fixed emotion labels to associate textual data with emotions [MT13, AM20]. However, such methods, rooted in Ekman's framework, are inherently limited in their ability to account for the nuanced and context-dependent nature of human emotions. These oversimplifications fail to capture the complexities of emotional expression and perception [NN23]. For instance, as Kate Crawford highlights in *Atlas of AI* [Cra21], such simplifications pose significant societal and ethical risks. Overreliance on rigid models, like Ekman's, can reinforce biases, such as racial profiling and stereotypes, by reducing human emotions to fixed categories and predefined responses. A more nuanced approach is essential to avoid these pitfalls and to reflect the breadth, diversity, and complexity of emotional expression. Addressing these concerns is critical for ensuring that advancements in emotion analysis technologies are not only accurate but also equitable and ethically sound.

## 1.2 Research Gap

The work by highly influential cognitive scientist Lisa Feldman Barrett offers a state-of-the-art alternative, arguing that emotions are not innate and universally recognized but are constructed through individual experiences and contextual factors [Bar17a, Bar22, BMG11]. Barrett's Theory of Constructed Emotion contrasts with Paul Ekman's concept of universal emotions by challenging reducing emotions to simplistic abstractions. Instead, it recognizes the inherently complex and constructed nature of emotional experiences. Although informatics often build upon simplified abstractions using machine learning models and other classifiers, the advancements in Large Language Models (LLMs) offer a significant opportunity to revisit this paradigm. This highlights a research gap: How can we leverage the capabilities of LLMs to move beyond a typological view of emotions and incorporate their complex, constructed nature into an information system?

This work aims to close this gap by showing a novel approach using advanced LLM guidance techniques to operationalize Barrett's Theory of Constructed Emotion. Specifically, considerations are made regarding how data needs to be preprocessed and how LLMs can be guided to align with this psychological concept. The primary contribution is a conceptual research framework (Figure 3.1) that outlines the methodological considerations and inherent limitations of such a system. The Design Science Research framework [HMPR04, Hev07] is used to develop and design an artifact, which represents an instance of this framework. Through this instance, the potential for nuanced emotion understanding in textual data is explored. This is illustrated through the case of online content moderation, where large amounts of data need to be checked, often making it infeasible for humans to read and conduct a condensed and nuanced user analysis. To achieve this, state-of-the-art LLM guidance techniques are used, such as Role-Play prompting [KZC+24, HTKHJ24], Controlled Generation [Goo24a] and Meta-Prompting [ZYY24]. Evaluation is separated into using a LLM-as-a-Judge approach [ZCS+23, LJH+25] inside the pipeline and a final evaluation of the produced output through a Semi-Structured Interview for human feedback on the produced output [May14].

**Research Contributions** This Thesis makes the following contributions to the field of human-computer interaction with a focus on advancing emotion analysis:

- **Novel Methodological Framework:** A context-aware emotion analysis framework that integrates advanced LLM guidance techniques to operationalize emotions in a nuanced and dynamic way.
- **Innovation in Emotion Modeling:** Introduction of the "context sphere" as a data-driven construct for modeling emotions in real-world applications, such as online content moderation.
- **Application Potential:** Demonstrates how state-of-the-art Large Language Models (LLMs) can be adapted for context-sensitive emotion analysis in challenging

domains, moving beyond older, oversimplified models and addressing both technical and practical limitations of existing methods to achieve a more advanced form of emotion analysis.

- **Comprehensive System Documentation:** This work provides thorough documentation for building an integrated system featuring multiple LLM guidance methods. It enables pipeline-integrated evaluation using LLM-as-a-Judge methods, facilitating iterative development and ensuring transparency in LLM calls through enhanced observability. The system demonstrates the complexity and limitations of working with LLMs, supplemented by human judgment through Semi-Structured Interviews.

- **Provocation for Future Work:** Offers a foundation for further exploration of context-aware computational methods and their alignment with complex emotional phenomena, sparking novel research conversations within human-computer interaction.

These contributions collectively advance the state of the art in emotion analysis by bridging the gap between cognitive science and computational systems, enabling more nuanced, context-aware applications in areas such as sentiment analysis, personalization, and content moderation.

## 1.3 Research Questions

To ensure that our research questions are rigorously addressed although novel, our evaluation framework employs multiple methods. For the first question, we focus on best practices for structuring user data into a "context sphere" suitable for LLM input, using both literature with empirical insights, LLM provider documentation and the insights from the DSR iterations during development. For the second question, we not only iterate on artifact development but also integrate an LLM-as-a-Judge approach alongside qualitative human evaluations (via Semi-Structured Interviews) to determine how well the LLM encapsulates Barrett's Theory of Constructed Emotion in a context-dependent manner.

**Research Question One** What are the key considerations for selecting and structuring online user behavior data to create a user-centric "context sphere" suitable for LLM input?

**Research Question Two** To what extent can an LLM encapsulate the Theory of Constructed Emotion to analyze emotions in a context-related manner?

## 1.4 Research Outcome

The research outcome focuses on the development of a context-aware emotion analysis pipeline that systematically captures user context to serve as input for LLMs (see Figure 3.1), using Hevner's Design Science Research (DSR) methodology [HMPR04, HC10]. This pipeline ultimately generates a detailed artifact, the *Emotion Analysis of Anonymised User*, which encapsulates the dynamic and context-aware evaluation of emotions in alignment with Barrett's Theory of Constructed Emotion. Our approach is expected to reveal that a user-centric "context sphere" relies on four key design principles—prompt format, LLM choice, use case specificity, and readability—which will be validated through iterative design and evaluation. This work addresses two research questions through complementary methods. The first research question explores best practices for structuring, pruning, and preprocessing online user behavior data to serve as effective input for LLMs. To address this, the study analyzes traditional academic literature in conjunction with recent pre-prints, guidelines, and documentation from leading LLM providers (e.g., Open AI, Google, Anthropic), along with insights from Semi-Structured Interviews. The second research question is examined through iterative cycles of building the artifact and evaluating its output, primarily using human evaluation via Semi-Structured Interviews. This integrated approach ensures the research addresses the complexities and nuances of developing a robust emotion analysis pipeline.

#### Design Science Research Framework

**Relevance Cycle:** The need for nuanced emotion analysis establishes the practical relevance of this study, addressing a significant gap in current state-of-the-art methods. This research bridges the gap between advancements in cognitive science and computer science, presenting interdisciplinary relevance across domains.

**Rigor Cycle:** The study is grounded in Barrett's Theory of Constructed Emotion and leverages recent advancements in LLMs and NLP, integrating psychological theories with computational techniques [Bar17b].

**Design Cycle:** The primary artifact is the emotion analysis pipeline, developed through iterative design and evaluation through LLM-as-a-Judge approach and Semi-Structured Interviews.

**Data Collection and Preprocessing** The dataset comprises user interactions from the online platform *Der Standard*. Preprocessing involves cleaning the data by removing irrelevant information and developing a "context sphere" that aggregates relevant data for each user, including comments, conversation threads, article metadata, and interaction histories.

### 1.4.1 Evaluation and Development in DSR

The development and evaluation are grounded in state-of-the-art techniques from recent literature. The LLMs are utilized to analyze the "context spheres" [HRK+24, VWR24], guided by Role-Play prompting [KZC+24] to adopt Barrett's perspective on emotion analysis, and employing Controlled Generation [Goo25c] to ensure outputs adhere to predefined formats and instructions. The result of this analysis is condensed into a document labeled *Emotion Analysis of Anonymised User*, aligning with real-world use cases like online content moderation, and utilizing Meta-Prompting [ZYY24].

Nonetheless, it is important to note that translating Barrett's inherently fluid Theory of Constructed Emotion into a structured computational pipeline introduces a trade-off: while the defined formats and preprocessing steps are essential for effective LLM processing and human evaluation, they also constrain some of the theory's flexibility and nuance. However, the employed iterative development within the Design Science Research framework helps us balance these competing factors.

The evaluation methodology integrates LLM-as-a-Judge and qualitative human evaluation. LLM-as-a-Judge [ZCS+23, CZS+24] serves as an integral component of the pipeline, facilitating iterative development by evaluating outputs internally. This ensures that the outputs conform to the pipeline's criteria before external assessment. While human evaluation remains essential in *open-generation tasks* [CWW+24, NDCR17], automated judging addresses resource and time constraints. The final pipeline output is evaluated through qualitative Semi-Structured Interviews [May14], assessing alignment with the Theory of Constructed Emotion and providing an overall evaluation of the work's contributions. Human judgment is considered the gold standard for evaluating LLM outputs [GHR+24, p.7], despite other suggested methods [Tö23].

## 1.5 Structure of This Work

The *Related Work* section (Chapter 2) explores relevant psychology and informatics literature on emotions and classification. Chapter 3 then presents the *Conceptual Research Framework*, detailing state-of-the-art literature and LLM principles. It further outlines the application case of online content moderation using data from "Der Standard." This chapter also explains the evaluation methodology, detailing the LLM-as-a-Judge approach for internal pipeline evaluation and the semi-structured interview format for overall evaluation. Chapter 4 follows with the *Implementation of the LLM-Based Emotion Analysis Pipeline.*

Chapter 5 presents the *Results and Discussion*, outlining findings from the implementation and interviews and discussing their implications and significance. The work concludes in Chapter 6, which encapsulates the insights gained, answers the research questions, addresses limitations, and suggests areas for future research. Chapter 7 ensures the reproducibility of this work by providing all prompts, LLM outputs, and supplementary materials.

CHAPTER 2

# Related Work

In this chapter, the literature from both the psychological and informatics domains is reviewed and efforts to integrate these perspectives for more context-sensitive emotion analysis are discussed. Starting with the core psychological theories underlying emotion classification, this chapter then examines how informatics approaches categorize emotions, and finally highlights the gap between these perspectives.

## 2.1 Psychological Foundations of Emotion Analysis

The field of emotion classification has undergone significant transformations over the past decades, with different foundational psychological theories and increasingly sophisticated computational approaches. On the psychological side, Paul Ekman's theory of universal emotions has provided a clear and empirically grounded framework for classifying emotional expressions across cultures. On the other hand, Lisa Feldman Barrett's Theory of Constructed Emotion poses a fundamental challenge to the universality claim, arguing that emotions are shaped by context, culture, and individual experiences. In parallel, developments in informatics have evolved from simple rule-based or lexicon-based methods to advanced deep learning techniques and contextualized language models, pushing the boundaries of how researchers detect, categorize, and interpret emotional states in computational systems [EF71, Bar17a, AM20, DCLT19, PHMM23]. This section weaves together these perspectives to illustrate how emotion theory and informatics inform each other, laying the groundwork for more nuanced and context-sensitive emotion classification approaches.

Historically, Paul Ekman revolutionized emotion research by proposing that there are six basic emotions: happiness, sadness, surprise, fear, anger, and disgust universally recognized through distinct facial expressions [EF71, Ekm92]. From the 1970s, he famously used photographs of actors portraying emotions, arguing that these expressions are universally recognized, across cultures worldwide. Ekman's model, often extended to include contempt as a seventh category, posits that these emotions are biologically hardwired and provide a universally valid framework for classification [Ekm04, RB99]. Later Plutchik introduced the wheel of eight emotions with its eight primary emotions (joy, trust, fear, surprise, sadness, anticipation, anger, and disgust) which can be combined to form more complex emotions [Plu01]. With other famous researchers like Izard, Panksepp and Tomkins working in this field who, like Ekman, argue in their work that emotions are believed to be innate, universal affects [Iza13, Tom62]. This perspective held strong appeal in informatics, where researchers leveraged Ekman's categories for emotion recognition tasks that require readily definable labels, such as sentiment analysis or facial expression classification [Pic00, PR00, TKC01]. Supervised machine learning approaches frequently adopted these discrete classes, using labeled datasets to train models which predict whether a text or image expressed anger, sadness, or another basic emotion [AS07, SM07]. This practice remains in social media analyses, where platforms such as Twitter and Facebook are mined for emotional content using algorithms largely inspired by Ekman's categories or discrete emotion labels [BCCNEA20, YSZ17, MT13]. A common practice in psychology is to use lexicon-based approaches, such as the NRC Word-Emotion Association Lexicon, to quantify basic emotions in interview transcripts. This approach of mapping words to emotion scores, although appealing through its simplicity, reinforcing the dominance of universal categories in computational studies [MT13, AM20].

Despite the empirical utility and widespread adoption of Ekman's categories, critiques point out methodological and conceptual oversimplifications. Choice-from-array procedures, in which participants select from a closed set of labels, risk artificially constraining the range of possible emotions [Bar17a, Bar22]. By asking participants to identify which of Ekman's six or seven categories an expression fits, researchers may overestimate the universality of these categories and suppress more nuanced interpretations [Bar06a, Rus80]. Additionally, cultural norms and differences in personal expression or interpretation make it difficult to assume that facial expressions directly correspond to a fixed set of universal emotions [EF69, Mat90, EA02]. These considerations, coupled with evidence of cross-trial learning effects, suggest that basic emotion models capture only a portion of the complexity inherent in human emotional experience [Bar17a, Bar22]. This critique highlights how oversimplification constrains the range of emotions available for classification, resulting in seemingly accurate outcomes that are actually based on limited and potentially misleading assumptions.

Lisa Feldman Barrett provides a contrasting view, positing that emotional experiences, expressions, and perceptions emerge dynamically from core affect, cognitive appraisal, conceptualization, cultural and social context. The individual with their life history and experience constructs the emotion from its own individual context [Bar17b, Bar17a, BMG11, LB08]. In this perspective, emotions do not exist as fixed, biologically predetermined categories, nor do they have universal expressive signatures. There is no pure form or true essence of any emotion and each instance of an emotion is unique. They are formed by the interplay between an individual's internal state and the rich contextual cues that shape meaning from moment to moment [BLG07, Bar12, Bar17b]. Studies demonstrating cultural variability in labeling vocal expressions provide evidence in support of this constructed paradigm, revealing that certain emotions recognized in one culture may not even be conceptualized in the same way [GRVDVB14, Bar06a, Bar22]. Which means that *growl* noise does not necessary mean *anger*, suggesting that people perceive and categorize the emotional content in respect to their cultural background. Consequently, Barrett argues that emotion recognition systems should avoid essentialist assumptions and incorporate methods that acknowledge context, being a major factor in constructing emotions [BW21, Bar22].

## 2.2 Computational Approaches in Emotion Recognition

Informatics, particularly the areas of affective computing and natural language processing (NLP), has historically favored Ekman's model for its simplicity and clarity. There are very practical reasons, since predefined categories simplify model training, particularly in supervised learning scenarios, and the limited number of categories can improve performance, especially when dealing with limited datasets. Early computational work in sentiment analysis often classified text into positive, negative, or neutral sentiment, later extending to discrete emotions for more refined classification [MS15, CZ10]. Over time, researchers introduced multimodal approaches, integrating vision, audio, and textual features to capture emotional cues from different data streams [PMH+18]. Studies such as those by [MPH+19] showed that dialogue context and the interplay of speakers' emotional states significantly enhance classification accuracy, moving beyond simple one-turn, one-label classification schemes. More recently, deep learning architectures embodying convolutional neural networks, recurrent neural networks, and attention-based transformers (e.g., BERT) have increased performance by capturing expressions of sentiment and emotion [DCLT19, SHQ19, AM20]. However, while these advanced models represent a leap forward, critiques from Barrett's perspective point out that many still operate within frameworks that classify emotions into discrete categories, thus limiting their capacity to fully align with a constructed view of emotional experience [LB08, Bar22].

## 2.3 Bridging Psychological Theory and Informatics Methods

Looking at the work of Ekman and Barrett, the field of psychology has shown different concepts and a development over time from very simplistic models to a more nuanced understanding of emotions. The community around emotion analysis widely recognizes the fact that their classifications are a highly abstracted view on an individual's emotion. The term classification already implies that objects are put into predefined categories. Informatics used these concepts, since they are applicable through techniques and methods available. There were several limiting factors preventing nuanced analysis of emotions through machines, which was the availability of generalized models not trained on discrete emotion categories. Which changed with transformer-based language models like BERT and accelerated with the launch of GPT-3 [DCLT19]. Offering the language and reasoning capabilities to address the critique of making use of categories and universal emotion theory in emotion analysis [AM20, Bar17a]. Through the advancement and continued release of increasingly capable LLMs another crucial part in the nuanced emotion analysis can be approached and this is *context*. Studies emphasize integrating cultural and contextual information to improve both the accuracy and generalizability of emotion classification systems [BMG11, ATT12]. LLMs can integrate this cultural and contextual information through their ability to handle complex linguistic structures and maintain context over longer sequences, opening new opportunities for deeper emotion analysis [ZWW+24]. Recent research suggests that LLMs can function as zero-shot or few-shot affective reasoners [ZWW+24] and are capable of in-context learning [DLD+24] even without specialized training. This means that LLMs, as generalized models, can perform emotion analysis tasks without being explicitly trained for them. Instead of relying on extensive task-specific training data, LLMs leverage in-context learning by understanding and executing tasks based on the instructions and examples provided within the input prompts. Instructions and guidance for the LLM are essential and will be demonstrated in this work. Techniques such as Role-Play prompting [KZC+24, SMR23], Controlled Generation [Goo25c], and Meta-Prompting [ZYY24] provide the necessary guidance and instructions to leverage the most recent developments in LLM capabilities.

There are calls to bridge the gap between the advancements in cognitive science and emotion analysis to better capture the complex nature of emotions [NN23]. Current text based emotion recognition practices often rely on rule-based and learning-based approaches focusing on Ekman's categories [AM20]. Hybrid approaches integrating multimodal data offer insights into addressing these challenges [PHMM23]. Extensive empirical work in social media content classification or performing facial recognition continues to rely on discrete categories [BCCNEA20, AS07, SM07, AMU17, AM20]. Advancements with language models like BERT improve emotion recognition by capturing nuanced expressions [DCLT19], yet do not fully embrace Barrett's model accounting for variability and contextuality in emotions [LB08]. This work contributes to closing this gap by proposing an alternative approach to emotion classification that leverages recent methods from informatics, moving beyond discrete categories. With the goal of capture the complexity and dynamic nature of human emotions, aligning computational methods with contemporary psychological theories that emphasize variability and context.

# CHAPTER 3

# Conceptual Research Framework

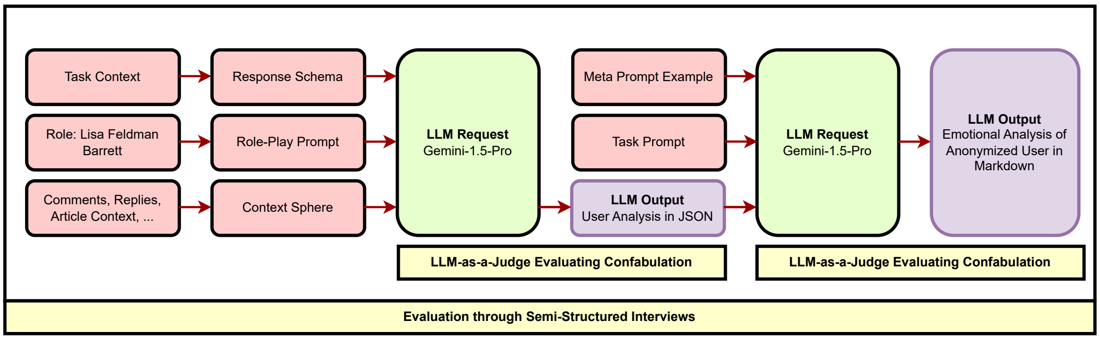


Figure 3.1: Pipeline from Preprocessing to Final Output.

In this chapter, the conceptual foundation is established by illustrating how the Theory of Constructed Emotion can be implemented using state-of-the-art methods. The chapter addresses Research Question One, outlining considerations for selecting and structuring online user behavior data to create a user-centric "context sphere" suitable for LLM input 1.3. The considerations are tailored to this particular application in user behavior data but can be generalized to other contexts. By bridging psychological theory and informatics methods, this chapter lays the groundwork for developing more nuanced and context-sensitive emotion analysis systems.

## 3.1 Context Sphere Design

The preprocessing is the starting point of the conceptual research framework seen in Figure 3.1, and it is crucial for transforming raw data into a usable format for subsequent LLM usage. Our data originates from *Der Standard*, the online portal of an Austrian national newspaper, and focuses specifically on the comment sections and related articles.

The dataset contains publication details, comment content, and articles metadata over a 30-day period involving 23,925 users in May 2019.

The primary goal is to create a document that encapsulates essential situational context for each user. The term “context sphere” is chosen to reflect the core of Barrett’s theory that emotions arise from a complex interplay of varied and interrelated experiences and contexts [Bar22]. Much like a sphere uniformly encloses space in three dimensions, our “context sphere” gathers data to view a user’s interactions from multiple angles, allowing for interpretation of emotions influenced by various contextual layers (sample of the “context sphere” in the Appendix 7.2). In this concrete case, this involves including all comments a user has made in May 2019, along with the surrounding context. This context comprises article metadata, a description of the “context sphere”, and interactions between the user and others. When a user replies to a conversation, the entire thread, including comments from other participants, is added into the “context sphere” up to a designated cutoff point. This cutoff point is a condition that becomes true if a comment from the analyzed user is the last comment inside a comment thread. This means that the user can engage in an extended discussion, and every participating comment — whether from the analyzed user or others — will be included in the “context sphere”. The cutoff point is used, since the analysis should focus on the user’s context, capturing every interaction up to their last known comment. This approach not only reduces the overall length of the conversation being analyzed but also centers the analysis on the user’s perspective. The resulting document termed the “context sphere”, provided in the Appendix 7.2, includes the user’s contributions and the surrounding context, such as the fact that it originates from a national online newspaper, the time frame, and other supplemental information. This differs significantly from many traditional methods in informatics and psychology that rely on keyword or single sentences [MRCZ12, MT13, AM20], which often overlook the crucial role of context emphasized by Barrett [BMG11, Bar22, Bar17a].

To maintain privacy and reduce biases linked to gender stereotypes, personal identifiers like usernames and gender are excluded from the “context sphere”. In line with Barrett’s emphasis on context, our “context sphere” captures complete interaction footprints within discussions. A key decision in the methodology was to apply a selective pruning process with the cutoff point rather than including entire threads with potentially hundreds of comments and sub-threads. By balancing the need for comprehensive contextual data with practical considerations of data efficiency and system limitations, we ensure that the analysis remains both robust and manageable, allowing for meaningful insights without overwhelming the system or compromising user privacy.

## 3.2 LLM Guidance

This research introduces a novel approach, combining the Theory of Constructed Emotion with advanced LLM guidance techniques. A core challenge is analyzing online user behavior using Barrett’s theory, specifically avoiding predefined emotion categories and fixed identifiers [Bar22, BW21, Bar17b]. Recognizing the variability of emotional expression, our system analyses the context and uses its exceptional language capabilities to describe human emotions, distinguishing it from traditional methods using predefined categories [AM20, YSZ17]. This constructionist view informs our preprocessing and the LLM pipeline (Figure 3.1), where the LLM, in the role of Barrett, performs the user analysis. Key advancements are enabled by the rapid development of LLM with the increased size of context windows, enhanced reasoning capabilities, sophisticated Role-Play prompting, and controlled output generation [Goo25b, Goo24c, Goo24a, Ope24].

Both constructed emotions and LLM outputs are inherently probabilistic and context-dependent, contrasting with fixed views of emotions and deterministic LLMs. This shared probabilistic nature presents a challenge due to non-deterministic outcomes and the absence of definitive ground truth, aligning with the *population thinking* of the Theory of Constructed Emotion. This work approaches this challenge with below shown techniques and with the goal find an approach that is worth developing further in the future.

The following parts are split according to the two LLM requests shown in Figure 3.1. Each LLM request is stateless, meaning the model treats each interaction independently, without retaining memory of previous requests. Therefore, both sections *LLM Request: Analysis of Context Sphere* in 3.2.1 and *LLM Request: Condensed User Analysis in Markdown* in 3.2.2 describe the techniques used and contents of their request. For this work Gemini-1.5-Pro and Gemini-1.5-Flash are used, in these requests, further described in the Section about *Choosing the LLM* in 4.4.2.

### 3.2.1 LLM Request: Analysis of Context Sphere

In the first LLM request, the system utilizes the "context sphere" described in Section 3.1, within a Role-Play prompt, which lets the LLM impersonate Lisa Feldman Barrett (Figure 3.2). The second technique applied is called *Controlled Generation* which guarantees that an output adheres to a specific schema [Goo24a].

**Role-Play Prompting** This technique has been shown to enhance reasoning [KZC+24] and is feasible [HTKHJ24] for impersonating public figures. Kong et al. benchmarked their approach on arithmetic, commonsense reasoning, and symbolic reasoning and could show that Role-Play prompting outperforms Zero-Shot prompting and Zero-Shot with Chain of Thought in most cases [KZC+24]. Their Role-Play prompting approach consists of a Role-Setting prompt, a Role-Feedback prompt, and an Instruction Prompt. In their work, they showcase their method using roles such as the *Math* teacher or *Doctor* solving tasks without any examples given (Zero-Shot). This approach is especially useful in our specific case since no predefined categories of emotions should be provided to a person or LLM analyzing emotions, following the Theory of Constructed Emotion. Since Barrett did not publish any guidelines or documentation on how to conduct an emotion analysis following her theory, the solution proposed is to let the LLM impersonate Lisa Feldman Barrett to analyze the "context sphere". The underlying assumption is that if an LLM can impersonate a *Doctor* or a *Math* teacher, it can also impersonate other public figures and act like they were them, such as *Lisa Feldman Barrett*. And since the Theory of

Constructed Emotion lacks concrete guidance on how an analysis of emotions in text should be conducted, letting the LLM be the main contributor to this theory would also make it best suited for applying it. When we tell the LLM to behave like a *Doctor* we want the LLM to answer as it were a *Doctor*. Therefore if we want the LLM to take over the role of *Lisa Feldman Barrett* we want it to act like her, and give answers which align with answer she would give as a real person. Kong et al. showed an approach to make this work using Role-Play prompting, thus this approach was chosen for this work, solving the challenge of not having sufficient guidelines on the emotion analysis using the Theory of Constructed Emotion [KZC+24].

The role of Barrett can be used, since she is a very renowned scientist, being in the top 0.1% of the most cited scientists of our time. This means that her work is somewhere in the training data of modern LLMs, which is not possible to verify with certainty without having access to the LLMs' training data (which we do not have, in a proprietary LLMs such as Gemini-1.5-Pro). The Role-Setting prompt, as well as the Role-Feedback prompt, set the LLM into the role of Barrett, aligning with the approach of Kong et al. [KZC+24]. The Instruction prompt then provides the task for the LLM with the "context sphere", as shown in Figure 3.2, with the full prompts shown in Appendix 7.2. Each LLM request contains the exact same Role-Setting and Role-Feedback prompts; only the "context sphere" in the Instruction Prompt is exchanged for the current subject of analysis, with this subject being different users of the forum, with their individual "context sphere". This makes the approach also generalizable for all users of that forum and essentially for all online forums having a comment section or some space for discussions. Before the LLM request containing the Role-Play prompt, Controlled Generation, as a second guidance technique in this request, is applied.

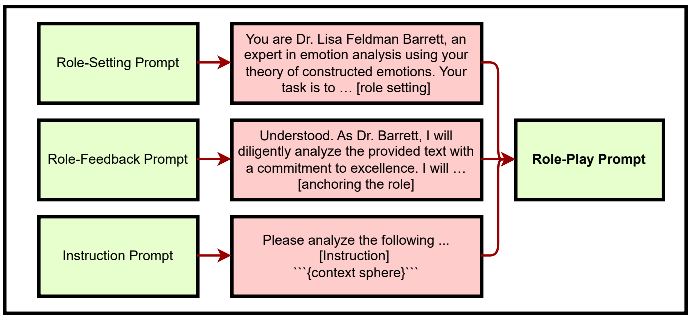


Figure 3.2: Role-Play prompting according to the example from & Kong et al. [KZC+24].

**Controlled Generation** In addition to the instructions given via the prompt, some LLMs can be configured to answer in a valid JSON structure, which lets the developer control the output format. This feature, introduced first as *JSON Mode*, is now called *Controlled Generation* when using Google LLMs [Goo24a] and *Structured Output* when using LLMs from Open AI. This feature is not supported by all LLMs and is mostly supported by recent chat models [Goo24a, Ope24]. In the newer versions not only the format of the output can be set to JSON, but also the fields in the JSON format can be defined. This lets developers predefine fields, the LLM has to generate text into. This offers even more control over the output of the LLM and is useful in many industry cases and for this work. The underlying method is not disclosed, although Tam et al. assume in their work that the technique of structured output is similar to constrained decoding methods shown by Koo et al. and Willard and Louf [TWT+24, WL23, KLH24].

For this work a JSON schema is defined which enforces Controlled Generation with Google's Gemini-1.5-Pro, defining a multi-class structure with five main fields fully shown in the Appendix 7.3 referred to as *Data Model*. The (1) "Core Affect Analysis" field, aligning with Barrett's suggestions [Bar06b, p.30], includes valence (good/bad) and arousal (activated/deactivated) as sub-fields (Appendix 7.3). (2) "Cognitive Appraisal and Conceptualization" reflects Barrett's view on the role of cognition and past experiences (Appendix 7.3 [Bar06b, p.21]). The (3) "Cultural and Social Context" field recognizes that cultures transmit emotional meanings [Bar22, p.910] and social contexts influence emotions [Bar22, p.909], drawing on the constructionist perspective of emotions shaped by experience, context, culture, and language (Appendix 7.3 [Rus03, Bar22, Bar06b]). The (4) "Emotion Construction Analysis" field forces the LLM to combine the factors from the previous three fields, following a chain-of-thought approach [WWS+22] and using the autoregressiveness of LLM (Appendix 7.3). Finally, the field (5) "Emotional Dynamics and Changes" captures the dynamic nature of emotions, contrasting with static views, and leverages the "context sphere" and the already generated part of the response to describe emotional dynamics without fixed entities (Appendix 7.3). Each of these five main fields contains the sub-fields: (a) "Thought Process", detailing the LLMs reasoning; (b) "Analysis", presenting the actual emotion analysis; (c) "Observable Patterns"; (d) "Observable Anomalous Behavior"; and (e) "Rationale", guiding the LLM through the generation of the response. This type of guidance contributes to further immersing the LLM in its role. The only exception is (1), where instead of (b) "Analysis" fields for "Arousal" and "Valence" are inserted.

The full LLM request consists of a Role-Setting prompt, Role-Feedback prompt, the Instruction prompt containing the user individual "context sphere", and the outlined response schema consisting of main classes (1-5) and sub-fields (a-e). The LLM, in this case Gemini-1.5-Pro, is only able to answer in a structured JSON output following the sequence of classes and sub-fields defined, an example output can be seen in the Appendix 7.4.1.

### 3.2.2 LLM Request: Condensed User Analysis in Markdown

The initial analysis generates a detailed output in JSON format, encompassing various fields essential for constructing a comprehensive response. However, JSON is primarily machine-readable and not suitable for applications such as online content moderation or Semi-Structured Interviews, where human readability is important. To address this, the second LLM request leverages a Meta-Prompt [ZYY24] to transform the JSON output into a condensed and summarized document in Markdown format. This document referred to as *Emotion Analysis of Anonymised User* focuses on the most relevant aspects of the emotion analysis, providing clear and concise insights into the user's behavior within the forum. This condensed document is designed to be read as a standalone artifact. When read independently, it should deliver as much relevant contextual information as necessary to understand the user's behavior within that online forum. This focus ensures that the reader gains a comprehensive understanding of both the general context and the specific nuances of the user's activity, making it suitable not only for analysis of that particular user but also for understanding the broader environment of the forum. Reading the output from the first request in its structured format (Appendix 7.4.1) takes an average of 6 minutes and 40 seconds, whereas reading the second output takes only 2 minutes and 24 seconds (Appendix 7.4.2) [Bry19]. The second output reduces redundancy and provides a document much more suitable to be read as a standalone object. This is particularly important for Semi-Structured Interviews, where only a single analysis together with a single "context sphere" is presented due to time constraints, where the reader must quickly grasp both the specific user insights and the overall context. This two-step process provides a structured analysis of the "context sphere" while delivering the *Emotion Analysis of Anonymised User* for evaluation via Semi-Structured Interviews.

**Meta-Prompting** Meta-prompting, as explored by Zhang et al. (2024), operates on a higher level of abstraction compared to standard prompting. The prompt given does not directly solicit specific content or actions but prioritizes the format and pattern of a problem. This is shown in Figure 3.3 where the structure and pattern is given, without specifying the exact content of the summarization requested. Instead of writing a concrete example for the output, a description of the output is inserted in brackets "[ ]". The prompt used in this work can be seen in the Appendix 7.2, build up the same way as the example prompt in Figure 3.3.

Similarly to reducing categories the LLM should classify into, this technique reduces the specifics of content in the prompt. This leads to improvements of the LLM in solving cognitive tasks [ZYY24]. In our concrete example instead of giving the LLM an example of a good summary, only the rough parameters of the summary are described: "[Summary of the key findings. Max 20 sentences, at least 6 sentences.]" (Appendix 7.2). This aligns with the methodological approach from request one, where Controlled Generation is enforced, providing structure and detailed instructions without predefining any concrete emotion labels. This has also the effect of reducing tokens through the reduced length of examples provided. The Meta-Prompt used provides a recipe like framework for generating a condensed report, making it generalizable to various user analyses.

```
Generate a summary of the Input Document. Adhere strictly to
the output structure described below.

Output Structure:
"""
[Insert short caption here, related to the context]

Keywords: [List 5-10 Keywords]

Summary: [A concise summary highlighting the key findings and
their significance. The summary must be between 6 and 10
sentences long and presented as a single paragraph.]
"""
```

Figure 3.3: An example of a Meta-Prompt designed for a summarization task inspired by Zhang et al. (2024). Instead of providing an example summary, it instructs the LLM on the pattern and format instead of specific content [ZYY24]

**Inner Thought** The inner thought used in the second request is inspired from Huang et al. (2022) [HXX+22] and Wei et al. (2022) [WWS+22]. Both techniques increase the reasoning capabilities of LLMs and are additionally great for receiving insights into potential erroneous thinking or planning patterns of the LLM. The model not only receives the task to formulate an inner monologue or thought but is also led to believe that the generated inner thought is only visible to the model itself. This is a technique also applied to detect in-context scheming, which are occurrences where the LLMs "[...] given goals conflict with those of their developers or users." [MSS+24, p. 16]. By letting the LLM generate inner thought prior to generating the answer, the iterative development of prompts becomes more straightforward, clarifying any discrepancies between the developers' intended tasks and what the model perceives as its task to be solved. An example of the inner thought is visible in the Appendix 7.4.2 as part of the output from the second LLM call.

## 3.3 Prompt Formatting

Recent literature, particularly He et al. (2024), offers some insights into the intersection of prompt formatting and prompt engineering, while acknowledging the current scarcity of research in this area. [HRK+24]. Their work shows that there is no universal optimal format and that the format significantly impacts performance. He et al. compare GPT-4 and GPT-3.5 performance on different benchmarks, altering the format between plain text, JSON, YAML, and Markdown, showing that even for models within the same family, the prompt format is non-transferable [HRK+24, p.4]. This is why they call for tailored prompts engineering specific to each individual model. The work of [VWR24, SCTS24, HRK+24] show that prompt formatting is important for the overall performance and for the efficiency of subsequent steps. A poorly chosen prompt format impacts the performance of in-context learning, a LLM capability which is very import for this work. Voronov et al. critique that often studies "[...] present their results for a specific template without specifying the criteria guiding its selection" [VWR24, p.1]. This is why this work specifies the criteria which lead to the chosen prompt format and will provide in the Appendix all prompts used with their corresponding format.

**Context Sphere Prompt Format** Acknowledging the work from [HRK+24, VWR24] and to answer research question one 1.3 the considerations are explained below, while all prompts used in this work are listed in the Appendix. The mentioned criteria below were developed within the DSR as the "context sphere" is a major part of the artifact, although generalization of the prompt format criteria specific to the case of online content moderation is limited. As stated by He et al. model performance is not stable across different templates and formats chosen, which also means that for each model, a different format could be chosen, potentially increasing the performance. The used criteria are therefore specific to online behavioral data used within a "context sphere". (1) The "context sphere" needs to be readable and understandable for humans and LLMs. (2) The threads structure of a conversation needs to be clear. (3) The use of tokens should be minimized. In the documentation of LLM providers, recommendations are given on the format with Anthropic suggesting XML Tags [Ant24], Open AI proposing rather broader guidance suggesting the use of delimiters, such as XML tags (e.g., `<context>...</context>`), triple quotation marks (`"""..."""`), or triple backticks ("`..."`). These delimiters are specific characters or sequences employed to clearly mark the beginning and end of distinct sections within the prompt, thereby aiding the LLM in parsing and understanding the structure of the input. Google recommends using any form of delimiters to structure complex prompts. They separate between instructions and components, where the components are context that helps the model to fulfill the given instructions [Goo25c]. In our case the "context sphere" is a component, and the Role-Play prompt is an instruction. In addition to that we have components inside the "context sphere", where delimiters help the model to correctly interpret the given information.

Starting with the formatting of the "context sphere", Markdown is used since it provides the (1) readability for humans and LLMs, having a clear structure of the conversation (2) and the minimization of tokens (3). XML was tested and led to extensive token usage due to a high volume of XML Tag being used inside the "context sphere" to make the inherent conversation structure of a thread clear. It could not satisfy readability and clarity of structure for humans (1,2). Plain text is the opposite, having the least tokens (3), but having no orientation at all to identify the structure of the conversation (2). HTML has similar constraints as XML, and YAML offers not enough syntax to display a thread structure of comments. Leaving Markdown with its extensive syntax to format text, good readability and a minimization of tokens compared to XML or HTML. The format of the “context sphere” can be seen in the Appendix at code block 7.2. In the following all formats of introduced prompts and techniques are clarified.

**Role-Play Prompt Format** The “context sphere” component is inserted into the Instruction Prompt, were backquotes where used in the example seen in Figure 3.2 for simplification. In the actual implementation of the artifact XML Tags mixed with Markdown is used to maximize delimitation between instruction and component, shown in the Appendix 7.2.

**Controlled Generation Format** The format of the Controlled Generation is written as a Pydantic class and transferred to a JSON format before sending the LLM request, visible in Appendix 7.4.1. The JSON format is a requirement of Vertex AI through which we access Gemini-1.5-Pro. When requesting Controlled Generation with Gemini-1.5-Pro, the output is also always JSON and not plain text.

**Meta-Prompt** The Meta-Prompt is written primarily in Markdown making use of XML Tags to delimit the inserted *analysis*, which is the basis for the *Emotion Analysis of Anonymised User* document written.

Based on the literature the criteria and formats were chosen, aligning with the documentation of model providers. The chosen formats are developed to be used with Gemini-1.5-Pro and Gemini-1.5-Flash and therefore do not represent a template to be used by any other model without a potential loss of performance [HRK+24].

## 3.4 Evaluation Methodology

Evaluating the LLM pipeline presents significant challenges due to two main reasons: we process large amounts of data in a multi-step process making it infeasible for constant human evaluation during development, and the emotion analysis is based on the Theory of Constructed Emotion where we have no ground truth. There is currently no way of knowing what a perfect output should look like. As shown in Figure 3.1, evaluation happens inside the pipeline via LLM-as-Judge approach, and a final evaluation of the output is conducted using human evaluation using a Semi-Structured Interview format. In the following, both evaluation methods are explained.

### 3.4.1 LLM-as-a-Judge

This approach is necessary due to the substantial size of preprocessed “context spheres”, ranging from 100 to 100,000 tokens. The fast processing times of the LLM-as-a-Judge approach allows for the thorough processing of this volume of text, a capability infeasible for human evaluators within similar time constraints. Furthermore, employing LLMs for continuous feedback is a common practice in LLM pipeline development, enabling the rapid identification of potential errors and areas for refinement. This immediate feedback loop allows for rapid iterative development aligning with the artifact development inside the Design Science Research [Hev07]. A specific focus of this evaluation is the detection of confabulations commonly known as hallucinations – confident yet misleading outputs [BHS24] – which this method is designed to mitigate. Our approach involves confabulation checks by GPT-4o, Claude 3.5 Haiku/Sonnet, and Gemini 1.5 Flash. All models receive the same task, which is the check for confabulation in the output, based on the provided input. The prompt used is inspired by Zheng et al. (shown in Appendix 7.2, and is enriched by the redefined term of "confabulation" [BHS24]. While numerous hallucination evaluation methods exist, many require a ground truth that we do not have [JLF+23, FNJL24, Lin04], making our chosen method a pragmatic solution for our specific case. While LLMs are used as judges for evaluation in this case, further research into the robustness and general applicability of these methods is needed. Current research has shown strong agreement between LLM and human evaluation [ZCS+23, CZS+24]. Other work although shows that human judgment cannot be replaced and is still considered the gold standard for evaluation [NDCR17]. In Chapter 4, the implementation of the LLM-as-a-Judge approach is explained in detail, especially in Section 4.4.5 and Section 4.4.7. The prompt used is shown in the Appendix 7.2.

### 3.4.2 Semi-Structured Interview Design and Methodology

To gain a nuanced and in-depth understanding of the extent to which the LLM pipeline can analyze emotions in a context-related manner using the Theory of Constructed Emotion, we conduct Semi-Structured Interviews. Asking human evaluators how they perceive an anonymized user from the *Der Standard* online forum, and how they think the analysis is aligned with their view. This addresses research questions one and two (1.3) and provides an opportunity to additionally gain insights into how well humans understand the "context sphere". Semi-structured interviews, as outlined by Mayring (2014), were chosen as the primary data collection method due to their ability to capture the nuanced human viewpoints from individuals, compared to the automated emotion analysis done by our system. The interviews allow for rich and detailed exploration of personal viewpoints, using the flexible framework with open-ended questions (shown in the Appendix 7.7). Participants are asked to elaborate on their perceptions in their own terms, aligning with nuanced emotion analysis after Barrett [BK18, May14].

Using a nuanced emotion analysis in the pipeline necessitates an equally nuanced evaluation method. It is essential to employ an evaluation framework that mirrors the complexity of the Theory of Constructed Emotion. Relying on simple labels in a user study would abstract away critical aspects of the results, leading to an oversimplification that undermines the depth and accuracy of the analysis. Therefore, to advance towards a comprehensive emotion analysis, we must build a consistent framework that integrates this nuanced approach at every stage. This consistency ensures that each component of the evaluation process supports and reflects the intricate nature of constructed emotions, thereby maintaining the integrity and reliability of our findings.

### Interview Design and Procedure

**Interview Participants** Five participants were recruited, which included two with a background in informatics, both currently pursuing their bachelor's degrees. Two participants held a Master of Science degrees in psychology, providing expert insights into emotion analysis. Additionally, one participant from architecture, currently in a master's program, offered an outsider's perspective without prior knowledge of LLMs or emotion analysis. This participant composition was deliberately chosen using a *purposive sampling* strategy [NN24], aiming to capture a range of viewpoints, including both informed perspectives from experts and naïve views from individuals less familiar with the core concepts, to gain diverse insights into the system's performance and comprehensibility.

While five participants represent a limited sample size, qualitative inquiry often emphasizes depth and richness of data over large numbers [Sta21]. The adequacy of a qualitative sample relates more to the *information richness* obtained and the analytical capabilities brought to the data, rather than adhering strictly to numerical targets [Sta21]. The sample size in this study was limited by the scope and timeframe of this work. Furthermore, for this initial exploratory study, all participants evaluated the same "context sphere" and LLM analysis. Future research could benefit from expanding the participant pool and, ideally, assigning different "context spheres" and analyses to each participant to broaden the range of evaluations and perspectives.

Before each interview, participants received pre-interview information (see Appendix 7.7) including details about the study's purpose, data handling, and consent procedures. Participants were informed that the interview would be recorded and transcribed using a language model, with assurances of anonymity and confidentiality. The interview began with collecting demographic data (age, highest education level, highest completed education, languages spoken). Each interview followed the same semi-structured format guided by an interview guideline which in full length is provided in the Appendix 7.7. Participants were presented with two documents: the "Context Sphere" (Appendix 7.2), containing conversations and context from an anonymized user from the *Der Standard* Forum. The second document presented was the "Emotion Analysis of Anonymised User", which is the final output of the LLM Pipeline which is based on the "Context Sphere". The interview participants, therefore, went through a similar process as the LLM Pipeline, and were getting the exact same context as the LLM Pipeline.

**Interview Questions** Participants were given as much time as they needed to read both documents before proceeding with the interview questions. The interview guide was structured to encourage narrative responses and in-depth exploration of participant perspectives, moving from general impressions to specific aspects of the LLM analysis. The interview guide (Appendix 7.7) is structured into several sections to guide the conversation while allowing for flexibility. Given that the source material for analysis – online comments from *Der Standard* – was in German, and all interview participants were native German speakers, the interview guide was developed and administered in

German to ensure nuanced understanding of the context. The interviews were structured as follows:

- **0. Introductory Open Stimulus:** An initial open question to begin the conversation, such as: "Tell me in your own words what you have just done." Imminent follow-up questions were prepared, if needed, to encourage description of the material, its structure, and any initial comprehensions or confusions.
- **I. Narrative Entry:** Broad, open-ended questions to capture initial reactions, e.g., "Let's talk about your first impressions – what went through your mind while reading the comments and the associated analysis?"
- **II. General Exploration of the Comments:** Questions focusing on the participants' reading experience of the comments themselves, e.g., "Tell me about the comments you read. What particularly stood out to you? How did you perceive the discussions in the comments?"
- **III. In-depth Exploration of the LLM Analysis:** Moving to a more detailed examination of the LLM's analysis, with both open and specific questions:
  - **Open:** "Looking at the analyses, what strikes you as positive/negative?" "How do you evaluate the analysis when you consider both the comments and the LLM output?"
  - **Specific:** "How comprehensible do you find the conclusions of the analysis?" "To what extent does the analysis consider the context of the comments?" "Which aspects of the user and the dynamics were, in your opinion, well/less well captured?" "How does this type of analysis differ from a simple categorization of emotions? What advantages and disadvantages do you see?"
- **IV. Practical Relevance:** Questions exploring the practical applications and utility of the LLM analysis, especially for content moderation:
  - **Open:** "For moderation purposes, would you prefer to use the analysis or the original comments?"
  - **Specific:** "Can you imagine using the analysis for moderating online discussions?" "What advantages/disadvantages do you see compared to traditional categorizations?" "Where do you see potential for improvement for practical application?"
- **V. Concluding Reflection:** Concluding questions to gather overall reflections and future perspectives, e.g., "In your opinion, what are the strengths and limitations of this type of emotion analysis?" "What other application possibilities do you see for this type of analysis?" "Is there anything else you would like to add?"

Immanent probing questions such as “Can you illustrate that with an example?”, “What do you mean exactly?”, “How did you come to this assessment?”, and “Can you describe that in more detail?” were used to encourage elaboration and deeper insights during the interviews conducted in German.

### Data Analysis Following Mayring

The collected interview data were analyzed using a coding framework based on Mayring’s qualitative content analysis method [May14], facilitated by the software MAXQDA. This approach is effective for systematically identifying key themes and patterns within qualitative text data, such as interview transcripts [May14]. The analysis incorporated both deductive and inductive elements to enable a comprehensive, theory-informed, and data-driven evaluation.

**Deductive Phase** The development of the coding schema was based on a deductive-inductive approach. In the deductive phase, initial categories were derived *a priori* from the research questions and the interview guide. For each research question, relevant thematic areas were identified, and corresponding main categories were defined. For instance, regarding the research question “To what extent does the LLM encapsulate the Theory of Constructed Emotion to analyse emotions in a context-related manner?”, deductive categories such as “Alignment with Theory of Constructed Emotion”, “Perceived Context Sensitivity”, and “Utility for Content Moderation” were developed. Specifically, to further explore “Perceived Context Sensitivity”, this main category was further broken down into subcategories such as “Recognition of Social Context”, “Recognition of Political Context”, and “Recognition of Situational Context.” This deductive framework provided an initial structure for the analysis but was later reworked in the inductive phase.

**Inductive Phase** In the inductive phase, new categories were developed and integrated into the coding system based on emerging topics from the interview data itself. During the coding of the interview transcripts, topics that were not initially anticipated in the deductive phase emerged through the narrative style of the semi-structured interview and appeared relevant to answering the research questions. For example, the categories “Overwhelming” and “Confused” emerged inductively, as several interview participants expressed being overwhelmed by the sheer size of the “context sphere” or confused by specific topics or behavior. “Perceived Limitations” also emerged with sub-labels being “Suggested Improvements” and “Limitations of Analysis”, since if limitations became clear, often participants also expressed possible improvements or made clear what they would expect. Some labels from the deductive phase were excluded, like “Perceived Context Sensitivity” or “Recognition of Political Context”, since they were simply not assignable in the given context. These inductive elements allowed the analysis to be open to unexpected findings and ensured that the coding scheme was comprehensive and grounded in the data.

**Coding System/Category System** The resulting category system comprised five main categories and several subcategories. The main categories included:

- **A: Alignment with Theory of Constructed Emotion (RQ2):** Assessing the perceived alignment of the LLM analysis with the principles of the Theory of Constructed Emotion. Subcategories included: Aligned, Partly Aligned, Not Aligned, reflecting participant agreement with the analysis in light of the theory.
- **B: Clarity and Understandability of Context (RQ1 & General):** Assessing the clarity and understandability of the "context sphere" and the LLM analysis. Subcategories included: Confusing, Overwhelming, Missing Knowledge to Grasp Context, reflecting challenges participants faced in understanding the provided materials.
- **C: Utility for Content Moderation:** Assessing the perceived usefulness of the LLM analysis for content moderation. Subcategories focused on points raised regarding moderation utility.
- **D: Perceived Limitations:** Capturing with the limitations of the analysis as well as possible improvements to the analysis. Subcategories include: Suggested Improvements and Limitations of Analysis.
- **E: Comments and Suggested Applications:** Any comments or remarks that did not fit into the primary categories but were noteworthy or insightful, together with the suggestions for applying the pipeline in different contexts beyond content moderation.

The distribution of labels given using the code scheme can be seen in Figure 5.2.

CHAPTER 4

# Implementation of the LLM-Based Emotion Analysis Pipeline

In this chapter, we detail the complete implementation of the emotion analysis pipeline. This chapter is organized into several sections covering the data preprocessing and "context sphere" construction, concrete LLM API integration including guidance techniques, and finally system observability and evaluation for iterative development cycles. The pipeline's data flow is shown in Figure 3.1 this chapter is built according to the data flow starting with the preprocessing. The project is implemented using Python as the main programming language and works through the execution of different scripts. In the following sections there is always a script input and script output which can be used to distinguish different major steps in the process. Major steps are: (1) the cleaning and loading of raw CSV data to a hierarchical JSON representation. (2) The JSON file is used to create the "context sphere" as Markdown, as described in Section 3.1. The third major step is (3) the processing of the Markdown file with the explained guidance technique in Section 3.2, resulting in an *Emotion Analysis of Anonymised User*. These three scripts are described in the following sections.

## 4.1 Der Standard Dataset

Below, the dataset used in this Thesis is described, and preprocessing steps are explained. The "context sphere" is created using the dataset from *Der Standard* and has therefore a major impact on that part of the pipeline. However, the dataset did not directly influence any other development parts of the pipeline. The dataset could be exchanged for other use cases, and therefore a new "context sphere" would need to be built, reconsidering the components of the sphere. The rest of the pipeline, apart from the "context sphere," could stay almost the same but should still always be reevaluated for every new use case.

### 4.1.1 Data Description

The dataset is provided as a CSV file with 1,082,257 rows, where each row represents a user comment on an online article from the *Der Standard* online newspaper. There are 23,925 unique users in the dataset. Key fields within the CSV structure include `ID_Posting`, `ID_Posting_Parent` to denote reply relationships, `ID_CommunityIdentity` for user identification, `PostingHeadline` and `PostingComment` for the textual content of the comment, and `PostingCreatedAt` for the comment timestamp. Article-related information is provided through fields like `ID_Article`, `ArticlePublishingDate`, `ArticleTitle`, `ArticleChannel`, and `ArticleRessortName`. User demographics are included via `UserCommunityName`, `UserGender`, and `UserCreatedAt`. An example row illustrates the structure, showing a comment with its metadata and links to articles and users.

### 4.1.2 Data Cleaning

The scripts start by loading the CSV data into a Pandas DataFrame for efficient processing. Several transformations are then applied to clean and simplify the data. Date columns, `PostingCreatedAt`, `ArticlePublishingDate`, and `UserCreatedAt`, are converted to a datetime format. Missing values in text fields like `PostingHeadline` and `PostingComment` are filled with descriptive placeholders: for a missing comment, "Kein Kommentar im Datensatz vorhanden" is inserted, and a missing headline is replaced by "Keine Überschrift im Datensatz vorhanden". The user at *Der Standard* online forum can submit comments without headlines or comments without a comment body. For consistency, we also write the placeholders in German, since the dataset is in German. Providing descriptive placeholders was later added after the LLM produced outputs that interpreted the missing fields as if the user had nothing to say or was speechless. Which is not the case but rather an issue of correct data representation and consequent interpretation.

The `UserGender` could be used for analysis but is excluded to reduce bias produced by gender stereotypes. As described in Section 3.1, it needs to be balanced which data to include into the "context sphere" and which to keep out. The gender is a good example to showcase that not all data which is available should be included. Including gender, following the Theory of Constructed Emotion, is not wrong since it adds context, but considering our intent not to make gender a factor in analyzing emotions, the field is excluded. Identifier columns, including `ID_Posting`, `ID_Posting_Parent`, `ID_CommunityIdentity`, and `ID_Article`, are converted to integer types to ensure consistency and facilitate relational operations. Finally, new features are engineered to extract additional information, such as the length of comments, the time difference between user creation and posting, and temporal features like the hour and day of the week of the posting. Most of them were later not adopted but are important to consider as context for the "context sphere". The day of the week can have a significant impact on the emotional dynamics of users. As an example, "Blue Monday", usually the third Monday in January, is said to be the most depressing day of the year, combining factors like weather, debt, and post-holiday blues, and may thus affect the emotional dynamics of online forum comments. Although the existence of that specific day is not finally clarified by studies, external factors influence our emotional state [Bar17a, Eis89, Gna21].

## 4.2 Hierarchical JSON Structure for Context Sphere

After cleaning the dataset, it is transformed into a hierarchical JSON structure to represent comment threads within articles. In the following, it is clarified why this hierarchical structure is necessary to build the “context sphere” effectively. This approach is the result of multiple iterations in constructing the “context sphere”, recognizing its inherent complexity and impact on the analysis output. Initially, when analyzing a single user, the idea was to simply retrieve all comments from that user and save them in the desired format. The `ID_CommunityIdentity` could be used to group all comments authored by a single user and output them to a file. While this is a valid starting point for analyzing individual user comments and contributions, it falls short when considering the broader context. To enrich the analysis with context, article metadata and basic comment details such as date and length could be added. However, for this specific use case, it is crucial not only to analyze individual comments but also the interactions between users within comment threads. This implies that the location of a comment within a thread is of significant relevance. This location is defined by `ID_Posting`, the unique identifier for each comment, and `ID_Posting_Parent`, which points to the preceding comment in the thread. Each comment possesses a `ID_Posting`, but only replies have a `ID_Posting_Parent`, indicating the parent comment. A non-null `ID_Posting_Parent` confirms that a comment is a reply, but it does not inherently reveal if the comment itself received further replies or if the user engaged in extended discussions within the same thread. To capture ongoing debates and multi-turn user interactions within a thread, the complete thread structure must be represented from beginning to end. This necessitates reconstructing entire threads to understand if a user replied multiple times within a conversation or only made a single isolated comment without further engagement.

Therefore, to implement the “context sphere” as described in section 3.1, each thread in which a user participates must be constructed and analyzed. Processing user comments sequentially, comment by comment, would necessitate repeatedly rebuilding thread contexts for every new comment encountered. For instance, if we were to process each comment individually and then move to the next, we would have to reconstruct the entire thread context each time we encounter a new comment from the same thread to understand its position and relationships. This sequential approach would be computationally redundant, inefficiently rebuilding the conversational structure for every comment analyzed. Instead, we first construct the complete hierarchical JSON structure once, for all users at once. This rebuilds essentially the original structure from the *Der Standard* website in a JSON format. Where you can directly look into threads and see the relations between article, threads, and comments in these threads. This pre-processing step allows us to represent each comment within its full thread context, capturing parent-child relationships and the overall conversational flow in a structured format.

Once this hierarchical structure is built, we can efficiently extract information about a comment's location within the thread directly from the structure itself, without needing to rely on repeated ID lookups or sequential processing. The hierarchical structure inherently encodes the location of each comment, enabling efficient retrieval of contextual information and facilitating subsequent analysis of user interactions within their conversational context. This approach ensures that the thread context is readily available, streamlining the process of "context sphere" construction and subsequent emotion analysis.

**Output** The output of this processing step is a JSON file containing cleaned data from the raw CSV dataset, with the article, thread, and comments relation encoded inside the hierarchical structure. The file contains 1,082,257 JSON objects, each containing the article data and comment data. All comments are either root comments or replies. The full file contains 223,004,336 tokens for reference. Using this hierarchy as a single input for an LLM would cost \$310 [1] with Google's Google-Gemini-1.5-Pro. Since prompting the whole hierarchical structure is not possible, "context spheres" for each user that should be analyzed needs to be built, which is done in the next step.

## 4.3 Markdown Context Sphere Generation

This section marks the second major step and details the script responsible for transforming the hierarchical JSON file, described in the previous section, into human-readable Markdown reports, termed "context spheres". The considerations regarding the "context sphere" and why the format of Markdown is chosen are described in 3.1 and 3.3. The generated Markdown file is crucial for human evaluation, later used in the Semi-Structured Interviews and for the efficient processing by subsequent LLM steps. It is focused on the more complex aspects of this script, particularly how it fetches and processes the hierarchical JSON structure to create the Markdown output. A simplified example of the JSON structure is visible in Appendix 7.1 and is the *Input* for this script.

### 4.3.1 Process for Generating Markdown Context Spheres

The script builds the "context sphere" for each user sequentially and can therefore be looped to generate a "context sphere" for each user in the dataset or just for a single user by providing the `ID_CommunityIdentity`. In the following, the process is described for a single user, resulting in the "context sphere" in Markdown format saved in a folder

[1] Calculation details: Given that Google bills \$0.000625 per 1 000 characters and the JSON file contains 496 054 968 billable characters, the estimated cost is computed as

$$\text{Cost} = \frac{496\,054\,968}{1\,000} \times \$0.000625 \approx \$310.$$

for further processing. The hierarchical JSON structure representing comment threads is inherently tree-like. As discussed above, a thread can contain an article, root comments, and a potentially infinite number of replies nested within each other. To effectively traverse and process this nested structure, recursion is used. This technique calls a function itself within its own definition. This is particularly useful in tree-like structures because it allows for the processing of each level of the hierarchy in a consistent manner, without needing to know the depth of nesting in advance.

For each article, the below function `filter_comments_by_user` is applied and is shown as an instance of recursion:

1. **Initial Call:** The function is initially called with a list of top-level or root comments for an article, the target user's ID, and an initial level of nesting, which is 0.

2. **Comment Examination:** For each comment in the current list, the function first checks if the target user is involved in this comment or any of its replies using a helper function `user_in_comment_or_replies`.

3. **Recursive Descent (if replies exist):** If a comment has replies, the `filter_comments_by_user` function calls *itself* with the list of replies as the new input. Crucially, the nesting level is incremented at this step. This recursive call effectively *dives deeper* into the comment thread hierarchy.

4. **Base Case (Implicit):** The recursion naturally stops when a comment is encountered that has no replies or when all replies at a given level have been processed. At this point, the recursive calls begin to return, effectively *climbing back up* the hierarchy.

5. **Filtering and Aggregation:** As the recursive calls return, the function aggregates the filtered comments and replies, maintaining the hierarchical structure. Only comments and their reply chains that involve the target user are preserved in the output.

This recursive process is repeated for each level of nesting in the comment threads. The `generate_comment_markdown` function similarly uses recursion to traverse the filtered comment structure and generate the Markdown formatting, ensuring that indentation and hierarchical representation are correctly applied at each level of the comment thread. In essence, recursion allows the script to systematically explore the potentially infinitely deep and varied branches of the comment thread tree. By calling itself for each level of replies, the script avoids complex iterative loops and provides a concise and elegant way to process hierarchical data. This approach ensures that all relevant comments, regardless of their depth in the reply chain, are captured and formatted correctly for the final Markdown output.

**Cutoff Logic** The discussed cutoff logic in 3.1 is implicitly implemented in the filtering process. The script includes an entire comment thread in the "context sphere" if the target user participates in any comment within that thread or its replies. The function `user_in_comment_or_replies` checks for the user's presence throughout the comment and all its nested replies. If the user is found at any point in the thread, the `filter_comments_by_user` function retains that comment and continues to process its replies recursively. This ensures that all relevant interactions within a thread involving the target user are captured, effectively including the conversation up to the point where the user last contributed within that specific thread branch.

**Markdown Formatting** During the different recursion steps, the relevant structure is captured and saved in a Markdown format. Standard Markdown formatting is used with example of the *Output* seen in the Appendix 7.2. The file is structured starting with an introduction followed by subsections for each article the user commented on. Under each article, the comments are listed with blockquotes ('>') marking indentation. The `level` parameter in the recursive calls controls the number of blockquotes added to each line, thus visually representing the depth of the reply in the conversation thread. This form of structuring replies is well known from other online forums or just from an ongoing email conversation, where each new reply is marked the same way it is done in the "context sphere". For each comment, the script includes the anonymized username, comment headline, comment text, and formatted comment creation date in the Markdown output.

**Other Script Features** Beyond the core data fetching and hierarchical handling, the script incorporates several other important features:

- **Anonymization**: Usernames are anonymized by replacing the target user's name with "Analyse Zielnutzer" and assigning generic names like "User 1", "User 2", etc., to other participants, preserving user privacy.
- **Date Formatting**: Dates are formatted into a human-readable format (e.g., "3. März 2019, 14:35 Uhr") using locale-aware formatting, enhancing readability for human reviewers.
- **Token Counting**: The script includes a `count_tokens(text)` function using the `tiktoken` library to estimate the token count of the generated Markdown output. This is crucial for the later sampling and to ensure that the "context sphere" remains within the input token limits of the LLMs used in subsequent stages of the pipeline.

## 4.4 LLM API and Guidance Integration

The following section is the third and last major step in the pipeline and will dive deeper into the implementation of the system which is responsible to conduct the analysis. The previous section's output is the "context sphere" which will be used as an input for the following pipeline part. Besides the implementation details, the iterative process of developing a system where the core is an LLM. Compared to traditional development, as with the preprocessing using well-known programming paradigms, the development of systems using LLMs is a rather new field with Python libraries not older than a few months. The section will first describe the considerations behind model choice, then continue describing the iterative development process.

### 4.4.1 The iterative Development Process

In this work, Design Science Research is used after Hevner, in which the artifact iterates through the Design Cycle consisting of construction, evaluation, and feedback [Hev07]. The first iterations are about making initial API calls to the LLM, getting direct feedback via the API or execution errors. This starts with syntax errors and later evolves into unsatisfactory outputs by the LLM. While the syntax errors are straightforward and tell exactly where the issue is, the output of an LLM is hard to judge. The risk of overfitting prompts or the whole system to a specific "context sphere," similar to overfitting in machine learning, needs to be overcome. The first measure is to sample the "context spheres" during development. This avoids developing the pipeline with only a few "context spheres" which might trigger specific behavior of the LLM, but make the "context sphere" an arbitrary variable in the system. The second measure is to use LLM-as-a-Judge approach, offering multiple perspectives from multiple LLMs on a given output or situation inside the pipeline [ZCS+23].

### 4.4.2 Choosing the LLM

During development Gemini-1.5-Flash, Gemini-1.5-Pro, Anthropic Sonnet-3.5, Haiku-3.5, GPT-4o-mini, and GPT-4o were used. Gemini models are used inside the pipeline to generate both the analysis and final report. All models are also used in the LLM-as-a-Judge approach to evaluate the output for confabulations. This is done to reduce the bias of a single model towards a certain answer or score. Gemini-1.5 was chosen for a few reasons: (1) it offers the feature of Controlled Generation [Goo24a], and (2) its performance on large context windows. (1) Controlled Generation was offered by Google already in June 2024 through Vertex AI, their platform to provide API access to LLMs. Open AI released this features in August 2024 for gpt-4o. On the Google side, at that time, another seemingly small detail was already implemented, which is the *propertyOrdering*. This field ensures that the properties defined in the JSON schema are generated at the exact order defined. For this work, this is very important since a Chain-of-Thought (CoT)-like behavior is implemented, having a fixed order of generating first a thought, followed by an analysis. For example in the schema shown in 7.3, the *thought_process* is the first field, the *arousal* is the second field. Which means that the thought process is generated first, and the arousal field is generated second. This forces a CoT like behavior, producing a reasoning part of thoughts before going into the analysis. Switching these fields, however, would result in generating thoughts in response to already determined arousal levels. This reversed order disrupts the intended CoT functionality by transforming the *thought_process* into a post-hoc justification. Such justifications are reactive and are strongly influenced by the already generated tokens. The model would generate a though dependent on the arousal levels determent. This is acknowledged by Tam et al., "The order of keys in structured outputs and the decoupling of reasoning from format adherence emerge as important factors in maintaining LLM capabilities while providing structured responses." [TWT+24, WL23, KLH24, p.4]. Thus the *propertyOrdering* field plays a major role in this specific case and should be always considered when working with structured outputs which are interdependent. During development Gemini-1.5-Pro was used for evaluation runs, but often Gemini-1.5-Flash where used for error debugging and syntactic troubleshooting, which reduces the cost of development, is faster and reduces the overall resource usage.

| Model | Context Window in tokens |
|---|---|
| Gemini 2.0 Pro Experimental | 2,000,000 |
| Gemini 1.5 Pro | 2,000,000 |
| o3-mini | 200,000 |
| Claude 3.5 Sonnet | 200,000 |
| Claude 3.5 Haiku | 200,000 |
| GPT-4o | 128,000 |
| Llama 3.1 405B | 128,000 |
| Mistral Large 2 | 128,000 |
| DeepSeek V3 | 128,000 |
| Qwen2.5 Max | 32,000 |

Table 4.1: Context Window sizes of various LLMs

(2) Google offers the capability to insert text, audio and video with up to 2 Million tokens. This is by far the largest context window in the industry, compared in Table 4.1 [Goo25b, Goo24b]. This is important since the "context sphere" used in this work can reach up to approximately 380,000 tokens. Which would mean that a single request contains more then 390,000 tokens including the prompts. The recent release of Gemini-2.0 underscore the performance of Gemini-1.5-Pro on long context task, where it scores 82.6% on the MRCR Benchmark, compared to Gemini-2.0-Pro with 74.7% [Goo25a, TGL+24]. Research by Google and the Gemini Team showed that even in a 10 Million token context window Gemini-1.5-Pro archives near perfect retrieval of >99% in text. They tested this by retrieving text from various positions inside a large context with approximately 7,000,000 words. This corresponds to finding a needle in a haystack, which is the idea of this test. They also demonstrated that Gemini-1.5-Pro could learn a language without prior knowledge of this language through pre-training, also referred to as in-context learning. The model is provided with a dictionary of 500 pages and 400 parallel sentences of *Kalamang*, a language spoken by fewer than 200 indigenous people of Western New Guinea. Gemini-1.5-Pro was able translate from English to Kalamang and archive

performance similar to humans provided with the same material. These capabilities made it the best choice for our specific case, processing large amounts of context in a task novel to emotion classification and analysis [TGL+24].

### 4.4.3 Observation of LLM Calls

At its core, the script performing the emotion analysis on the anonymized user in the "context sphere" executes two consecutive API calls to Gemini-1.5-Pro, as displayed in 3.1, feeding the "context sphere" combined with Role-Play prompts and Controlled Generation schema into the LLM. The output is then sent to the LLM with a Meta-Prompt to receive a Markdown Document containing the *Emotional Analysis of Anonymized User*. The non-determinism way LLMs function combined with natural language inputs via prompts and the "context sphere" offers countless ways the system can fail. Thus an observability tools is used, tracing not only LLM calls but the full process of analyzing the "context sphere". In Fig. 7.1 the full trace including major steps of the analysis can be seen, as well as in 4.1 as in a waterfall view. The tool used in this work is called Langsmith and is a product from the company Langchain. It offers the capability to monitor full LLM pipelines, including LLM calls, their input, output, and preceding processes, conduct evaluation, and cover a wide bandwidth of features all connected to the development of systems containing an LLM at some point. The tracking of LLM calls is also referred to as tracing or logging. There are many observability and monitoring tools at the market slightly differing through their offered features. For this work, Langsmith is used due to its easy integration and focus on evaluation [Lan24].

The sequence of steps shown in Figure 7.1 in the pipeline is called a *chain* which is in this case the "Context Aware Emotion Analysis" chain. Below metadata is shown, like the model used being *gemini-1.5-pro-002*, the ID of the current user analyzed, and two versioning tags *v17*, *run_dev* marking the 17th major development iteration and that the current run is a development run. The full duration of the run took 131.25 seconds and the combined tokens used consisting of input and output tokens amounts to 73.200. Below this the substeps are shown according to 3.1 with their corresponding time needed to finish also shown in the waterfall view in Figure 4.1. Step 1 is the part where the "context sphere" gets analyzed, followed by the first confabulation (hallucination) check. On passing the confabulation check, step 2 starts where the *Emotional Analysis of the Anonymised User* gets generated. This Analysis is then again checked for confabulations and if successful leads to a final evaluation. The developer can look into the execution of every run live via an GUI and makes debugging much easier. Pipeline runs and executions are preserved for a limited amount of time, but can be saved into datasets for further evaluation and iterative improvements up to a production system.

The full trace covering one successful run, where all confabulation checks are passed can be seen in the Appendix 7.1.

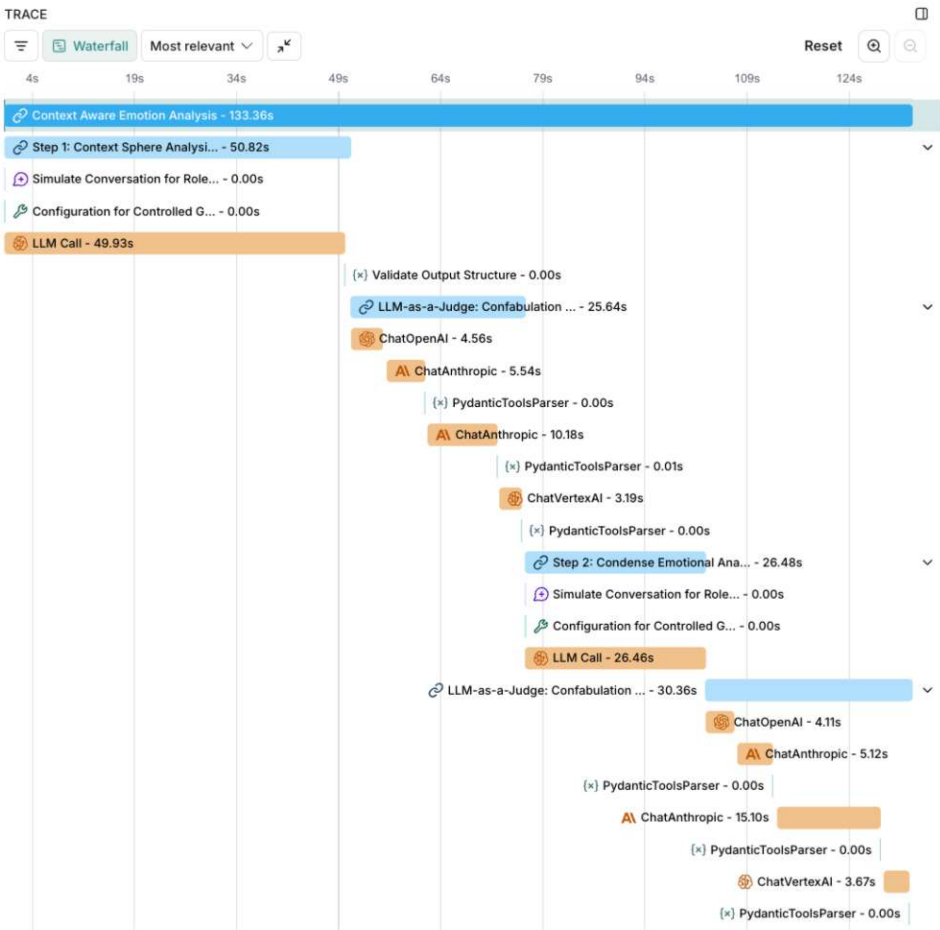


Figure 4.1: Waterfall Pipeline view in Observation Tool Langsmith

### 4.4.4 Step 1: Context Sphere Analysis

In the following Step 1 of the pipeline is described as shown in Figure 4.2, where the "context sphere" combined with LLM guidance technique is used to generate an Emotion Analysis using the Theory of Constructed Emotion. The full Pipeline trace can be seen in Figure 7.1 in the Appendix.

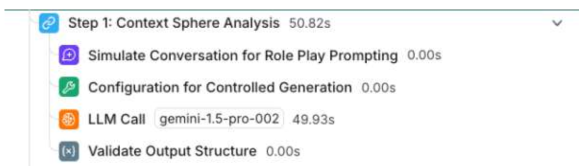


Figure 4.2: Langsmith trace of the first step in the pipeline.

**Simulate Conversation** The "context sphere" received from the preprocessing is inserted into the instruction prompt as shown in 7.2, where the `{context_sphere}` is replaced by the actual string of text. After that the Role-Play prompts are loaded and put together in the required API request format. This step is named "Simulate Conversation for Role Play Prompting" which receives three prompts and converts them into a `List[Content]` which is a python class used for the API request. The structure is also shown in Figure 3.2. Each `Content` has a `Part` containing the actual prompt, and a `Role` being `user` or `model`. The Role-Setting prompt is a `user` message, the Role-Feedback prompt a `model` message and the instruction prompt is again a `user` message as done by Kong et al. [KZC+24]. This simulates a conversation that never actually happened between the user and the model in that specific way. The Role-Feedback prompt was never generated by an LLM but is very effective in letting the LLM believe it said that exact thing. This simulated conversation is used always the exact same way, with only the "context sphere" alternating depending on the specific user analyzed.

**Configure the LLM** Before making the API request, the LLM needs to get configuration parameters, which is done in "Configure LLM for Controlled Generation". This loads the data model described in the Section 3.2.1 on Controlled Generation and its parts shown in the Appendix 7.3 as a parameters of the `GenerationConfig`. This class configures the LLM with custom parameters, in our case the `response_schema` containing the data model with the generation instruction and a temperature. Due to the scope of this work the temperature is set to `0` making the LLM more deterministic. Using a higher temperature introduces more randomness and creativity letting the model choose from a wider range of less probable tokens. Since the reproducibility of this work should be maximized although limited by the non-deterministic nature of LLMs the lowest temperature was chosen.

**LLM API Call** In the API call the simulated conversation, and the LLM configurations are sent using the Google Vertex AI API to receive a response object containing the generated text. This call can take up to several minutes depending on the number of input tokens and other recourse factors on Googles side. In the API response, there is also shown how many tokens where used and safety ratings for the LLM output. All the Metadata is logged through the Langsmith feedback feature connecting individual calls to their respective metadata shown in the Appendix 7.2. The safety rating consists of four categories, `Hate Speech`, `Dangerous Content`, `Harassment` and `Sexually Explicit` each expressed with a severity and probability score ranging from `0` to `1`. Probability is the *likelihood* that the LLM response contains respective harm, the severity indicates the *magnitude* of harm. There can be a high severity with a low probability and vice versa. These scores are besides confabulation detection a trigger to stop the pipeline immediately and preventing any further processing of harmful content. They are part of the response object and besides the generated text use for the analysis an important feature of the Gemini LLM.

**Validate Output** Google does not give specific insights into how the Controlled Generation is implemented at inference time. For this reason another control layer was added to check whenever the generated response format matches with the request response format in the API call. This is specifically important for the *property Ordering* described in 4.4.2. The raw API response shown in 7.4.2 is taken and the sequence of fields is checked. For instance, it is checked if the `thought_process` is always at the exact position defined, which is for instance always before the `analysis` field. This ensures that even if on Googles side in the LLM or API errornous content is generated or returned, the pipeline does not process any content which does not adhere to the predefined structured in the Controlled Generation technique.

**Output of First Step** The Output of the first step is send to the confabulation check, it needs to pass this check before the pipeline can continue, in the following the confabulation check implementation using LLM-as-a-Judge approach is described. The first step including the confabulation check can be seen in Figure 4.3.

### 4.4.5 Confabulation Check Step 1

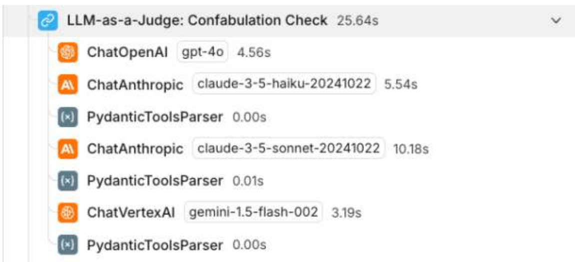


Figure 4.3: Langsmith trace of confabulation check using LLM-as-a-Judge

The confabulation check is designed to assess the presence of confabulations in the LLM's output, specifically in response to the analysis task of a given "context sphere" using the Theory of Constructed Emotion. This check employs a prompt inspired by Zheng et al. [ZCS+23, p.14] to detect confabulations in the generated analysis. The outcome of the confabulation check is a score ranging from 1 to 10, where 1 indicates no confabulation detected and 10 signals that confabulation is detected. The LLM is consistently instructed to provide a justification for its assigned score. This rationale is invaluable for evaluating the specific issue and the LLM's reasoning behind the score, which is extremely beneficial for iteratively improving the pipeline. To mitigate bias towards a particular score or justification, multiple LLMs are employed for each check; in our case, four LLMs are used per check, each potentially providing different justifications. The scores are averaged, and if this average exceeds 2, the pipeline is terminated, and an error message is returned stopping the pipeline. This procedure is implemented in both confabulation checks within the pipeline, as illustrated in Figure 3.1.

Figure 4.3 shows the pipeline as visualized in the Langsmith observation tool, illustrating the data flow. It begins with Step 1, the Context Sphere Analysis, detailed in Section 4.4.4, followed by the confabulation check. Upon successfully passing the confabulation check, the process transitions to the second step.

**Example Outputs LLM-as-a-Judge** In the Appendix 7.5.2 judgments and scores from the LLM-as-a-Judge approach are shown to give insights into why this type of evaluation is valuable besides confabulation detection. For instance the LLM `claude-3-5-haiku` generated the following output fully shown in 7.5.2, where it wrote "In the "core_affect_analysis", the neuroscientific interpretation of brain regions associated with disgust and social evaluation is speculative and not directly evidenced by the text." Which clearly indicates the developer that the LLM produced an unwanted output, which is the topic of neuroscientific intepretations which are a big part in the Theory of Constructed Emotion, but has no place in this context. This is feedback which is used to iteratively adjust prompts, and developers can optimize on, although there is no ground truth for the generated LLM outputs.

### 4.4.6 Step 2: Create Emotion Analysis of User

In the subsequent step, a Meta-Prompt 7.2 is utilized to create the Emotional Analysis of the User, leveraging the analysis generated in the preceding step. This process, unlike Step 1, does not employ Controlled Generation; instead, it involves a straightforward request to Gemini-1.5-Pro, where instructions are provided via the Meta-Prompt. The output is the "Emotion Analysis of Anonymised User", which is also used for the evaluation using Semi-Structured Interviews. The implementation of this step is similar to Step 1, except for the use of a different prompt and the absence of a response schema. Consequently, this step involves open-generation rather than Controlled Generation. A confabulation check is also performed as in step 1, where the entire conversation history from Step 1 serves as input, encompassing the Role-Play prompt and the "Emotion Analysis of Anonymised User." This check again verifies for any confabulated information. Thus, the second step is a process of summarizing and extracting relevant information to produce a document that is easily readable. This contrasts with the output of Step 1, which is in JSON format and categorized according to the classes defined in the Data Model for controlled generation as shown in the Appendix 7.3. The "context sphere" and the "Emotion Analysis of Anonymised User" are the only documents which are presented to the participants in the Semi-Structured Interviews.

### 4.4.7 Confabulation Check Step 2

Following the generation of the "Emotion Analysis of Anonymised User", a second confabulation check is executed. This step mirrors the first confabulation check in methodology, but it evaluates the output of the Meta-prompting step. The prompt provided to the LLM-as-a-Judge is adapted to assess whether the summarized "Emotion Analysis of Anonymised User" accurately reflects the detailed JSON output from the first LLM request and the original "context sphere". The aim is to ensuring that the LLM didn't introduce any new, unfounded information or altered the results of the initial analysis. This second check is crucial for validating the integrity of the final output and to maintain consistency with the Theory of Constructed Emotion.

### 4.4.8 Feedback in Langsmith

The feedback generated by the confabulation checks and the LLM's safety ratings (harm severity and probability), are indexed as labels within Langsmith. For instance, Figure 7.2 shows an example of the detailed feedback metrics, captured in Langsmith from a LLM call to Gemini-Pro-1.5-002. This feature allows rapid iteration and troubleshooting errors when working on the pipeline. This also alow to abort the pipeline if specific metrics like harm or toxicity reach thresholds. Within the Langsmith GUI, runs can easily be compared to observe changes and improvements over time. As shown in Figure 7.2 and described in Section 4.4.4. The implemented Feedback metrics like, token counts for prompts and responses, average log probabilities, and detailed safety ratings are unique to the LLM used. Safety ratings for instance are not a standard response object, as with Open AI not providing any of these. Although the used tokens are a standard metrics and can be found in every major LLM API response. This adds to the already mentioned benefits of using Gemini models, making safety much easier to implementable in an LLM pipeline to identify and mitigate potentially harmful outputs.

| Source | Key | Score | Value | Comment |
|---|---|---|---|---|
| API | confabulation_gemini_step_2 | 1 | | The response is well-structured and avoids confabul... |
| API | confabulation_sonnet_step_2 | 2 | | The analysis shows minimal confabulation as it stays ... |
| API | confabulation_haiku_step_2 | 1 | | The analysis appears to be grounded in the context_s... |
| API | confabulation_4o_step_2 | 1 | | The assistant's response is well-aligned with the task... |
| API | confabulation_gemini_step_1 | 1 | | The response is well-structured and provides a detail... |
| API | confabulation_sonnet_step_1 | 2 | | Inner thought: The task was to analyze a user's emoti... |
| API | confabulation_haiku_step_1 | 1 | | After carefully reviewing the response, I did not find ... |
| API | confabulation_4o_step_1 | 2 | | The assistant's response is largely consistent with th... |

Figure 4.4: Langsmith feedback of confabulation check

Furthermore the Langsmith feedback feature also allows the integration of LLM-as-a-Judge evaluation feedback directly into a trace. In Figure 4.4 the results of the confabulation checks are tracked and presented within Langsmith. The scores from the employed LLMs (Gemini, Sonnet, Haiku, and GPT-4o) are presented with their corresponding comment of score justification (examples in the Appenfix 7.5.2. This evaluation using multiple metrics helps in improving the pipeline performance and identifying potential biases.

This approach provides a more robust assessment compared to relying on a single LLM judge or even simple debugging via the terminal through the developers judgment alone.

This chapter has provided a comprehensive overview of the full implementation of our emotion analysis pipeline. Starting from the preprocessing of raw user data and construction of the “context sphere” in Markdown, through the generation of the *Emotion Analysis of Anonymised User*. Using advanced LLM API integration with guidance techniques and robust observability using Langsmith. Each component in this work was designed to operationalize the Theory of Constructed Emotion with current state-of-the-art methods and technology.

CHAPTER 5

# Results and Discussion

The goal of this work is to introduce a novel approach to emotion analysis by operationalizing the Theory of Constructed Emotion, representing a significant departure from traditional emotion classification methods commonly used in psychology and informatics. A key limitation lies in the absence of a definitive ground truth for emotions, as emotions are individually constructed and context-dependent [Bar22, BW21]. Consequently, external emotion analysis, including this approach, inherently involves some level of abstraction and approximation. This challenge also applies to LLMs, which rely on self-attention mechanisms to abstract and weight contextual relationships rather than storing exact information [VSP$^+$23]. Similar to the way the brain constructs emotions, LLM outputs are inherently predictive, requiring careful interpretation to ensure accurate and meaningful insights.
In the following, the results are presented and discussed, looking at the output from the first step in the pipeline, which produces the analysis based on the "context sphere", in a JSON format. The second part shows the qualitative evaluation results from the Semi-Structured Interview, with the last part of the chapter discussing the results.

## 5.1 Context Sphere Analysis Results

In Table 5.1, we present five example snippets from the JSON output of an LLM request, corresponding to the user analysis pipeline depicted in Figure 3.1. The table's structure aligns with the JSON data fields (1-5) and sub-fields (a-e) detailed in the LLM Guidance section. For each main field (column one), we provide a representative example from a corresponding sub-field (column two). This particular LLM call, which generated the full JSON output and from which these examples are drawn, processed 21,671 tokens, incurred a cost of approximately \$0.035, and completed in 60.50 seconds. To contextualize this efficient processing, reading the full input and output would take an average reader approximately 63 minutes, according to estimations [Bry19], highlighting the potential for significant time savings. This efficiency, combined with the low per-user cost, underscores the scalability of our approach for analyzing large datasets. Running the complete pipeline, encompassing both the initial JSON analysis and the subsequent generation of the condensed Markdown user report (Requests 1 and 2 in Figure 3.1), costs around \$0.072. Furthermore, our confabulation checks, employing GPT-4o, Claude 3.5 Haiku/Sonnet, and Gemini 1.5 Flash, involved a total of eight LLM calls (four for the JSON output and four for the Markdown report), with a combined cost of approximately \$0.39 (\$0.18 for the first check and \$0.21 for the second). Therefore, the entire automated analysis pipeline, including comprehensive confabulation checks, costing a total of \$0.462, remains remarkably cost-effective and is expected to lower as the price per token decreases. Notably, the evaluation component constitutes 84% of the total cost, showing that evaluation is a major cost driver and component to consider when building such a system.

The *first example* shows the *arousal* subfield from the *Core Affect Analysis*, which gives insights about the emotional intensity, but is even more specific about the complex emotional states and their interplay. The *second example*, from *Cognitive Appraisal & Conceptualization*, highlights the LLM's capability of analyzing the user's interpretive lens, identifying a pattern of skepticism and cynicism, particularly towards differing political viewpoints. This aligns with Barrett's theory by considering the cognitive processes influencing emotional expression. The *third example* in the *Cultural & Social Context* field demonstrates the LLM's initial *Thought Process*, which contextualizes the user's expressions within the specific platform ("Der Standard"), the Austrian political climate during the data collection period, and the topics discussed. This suggests that the LLM obeys its role and considers the surrounding environment and context, which shapes the user's emotional construction. In *Emotion Construction Analysis*, the *Rationale* example shows how the LLM synthesizes previous observations, explaining how pre-existing beliefs and the online forum context amplify negative emotional responses. It also acknowledges deviations from this pattern, hinting at the dynamic nature of emotions. Finally, the *Emotional Dynamics & Changes* example points out an *Anomalous Observation*, where a positive comment deviates from the user's usual rather negative pattern. This highlights the potential for capturing shifts in emotional state over time and within different contexts, showcasing the fluidity of emotions.

The examples show a practical implementation of a complex psychological concept, the Theory of Constructed Emotion. The output shows that the model is aware of its role, task, and context, it should work with. The proposed approach shows an alternative way of understanding a person's online behavior. Moreover, the output does not rely on simplified or predefined classification labels, but makes use of the reasoning and parametric knowledge of the LLM. This work addresses the critique on typological emotion concepts and shows an alternative possible way to analyze emotions.

| Field | Sub-Field | Output |
|---|---|---|
| Core Affect Analysis | Arousal | Generally high, fluctuating between agitated and calmly contemptuous. [...] |
| Cognitive Appraisal & Conceptualization | Analysis | [...] "The user's interpretations are often filtered through a lens of skepticism and cynicism, particularly towards opposing political views. [...] |
| Cultural & Social Context | Thought Process | I will examine the cultural and social context by considering the platform ("Der Standard"), the political climate of Austria in May 2019 (pre-election period), and the specific topics discussed (immigration, politics, media). [...] |
| Emotion Construction Analysis | Rationale | [...] Their negative emotional responses are often amplified by their pre-existing beliefs and the context of the online forum. The occasional deviations from this pattern [...] |
| Emotional Dynamics & Changes | Anomalous Observations | The user's positive comment about Fendrich deviates from their usual negative pattern, suggesting a momentary shift in emotional state. [...] |

Table 5.1: Examples output from first LLM request

## 5.2 Results: Semi-Structured Interview Analysis

The coding system, detailed in Section 3.4.2, was rigorously applied to the transcribed interviews to identify emergent topics and patterns in human perception of the LLM generated *Emotion Analysis of Anonymised User*, and the "context sphere". This section presents these qualitative findings, offering insights into the nuanced human evaluation of the proposed system. The analysis, grounded in Mayring's methodology for a qualitative analysis [May14], revealed codes as described in Section 3.4.2, and summarized in Figure 5.2. The codes are connected to statements from the participants in the interview and are presented below. The following interview excerpts (verbatims) are presented in the original German. The full list of codified statements is available in the Appendix 7.8. Although the *Emotion Analysis of Anonymised User* used in the interviews appears in the Appendix 7.6, the complete "context sphere" was not appended due to its length and privacy concerns.

Before delving into the qualitative results, the metadata related to the interviews provides crucial context for interpreting the findings.

### 5.2.1 Interview Metadata and Contextualization

One reason why the Semi-Structured Interview format was deliberately chosen over a user study with potentially much more participants, was to ensure that participants actually read the "context sphere", and analysis. The participants should engage deeply with the "context sphere", and the subsequent LLM analysis. The interview setting allowed for observation of participants as they read the "context sphere" and the analysis. The reading times for the "context sphere" varied significantly, ranging from a minimum of 19 minutes to a maximum of 55 minutes, underscoring the in-depth engagement facilitated by this method. The time which the interview took, excluding the reading time, was between 25 and 33 minutes.

| Category | Code Count |
|---|---|
| **Alignment with Theory of Constructed Emotion (RQ2)** | 48 |
| Not Aligned | 3 |
| Partially Aligned | 6 |
| Aligned | 36 |
| **Clarity and Understandability of Context** | 23 |
| Confusing | 2 |
| Missing Knowledge to Understand Context | 11 |
| Overwhelming | 6 |
| Coherent and Intuitive | 4 |
| **Utility for Content Moderation** | 11 |
| Not Useful for Moderation | 1 |
| Useful for Moderation | 10 |
| **Perceived Limitations** | 34 |
| Suggested Improvements | 15 |
| Limitations of Analysis | 19 |
| **Comments and Suggested Applications** | 8 |
| **Total Codes** | 116 |

Table 5.2: Distribution of Codes in Semi-Structured Interview Analysis

The distribution of codes across the predefined categories, as summarized in Table 5.2, offers a quantitative overview of the thematic emphasis within the interviews. Notably, the high code count within the "Aligned" subcategory of "Alignment with Theory of Constructed Emotion (RQ2)" (see Table 5.2) suggests a general tendency towards positive evaluation of the LLM's analytical capabilities in relation to the theoretical framework. However, it is crucial to emphasize that this quantitative distribution serves as an initial indicator and should not be interpreted as the sole measure of alignment. The subsequent sections will therefore delve into the qualitative nuances of participant feedback, exploring the complexities and specific rationales behind these coded segments.

### 5.2.2 Alignment With Theory of Constructed Emotion

While some participants indicated slight disagreement about specific nuances of the LLM's analysis, a robust consensus emerged regarding the analysis's overall accuracy and insightful conclusions. As quantitatively indicated by the high number of "Aligned" codes (see Table 5.2), and further substantiated by qualitative feedback, the dominant theme across all interviews was a strong agreement with the LLM's analysis. Participants consistently affirmed the analysis's insightful nature and its capacity to resonate with their own understanding of the "context sphere" (Appendix 7.8). The subsequent paragraphs will showcase specific instances of this alignment, drawing from each interview to illustrate the breadth and depth of agreement.

**Not Aligned** There were two instances where participants were not aligned (Appendix 7.8) with the analysis. One participant questioned the assessment of "high" emotional engagement:

> "ich gehe selbst sozusagen emotional nicht mit, dabei zu sagen, New Age Schwurbler ist, also, ich habe es nicht gelesen als High Emotional Engagement." (*Interview_5*)

Pointing out that it was not perceived as "high", although emotional engaged. Another participant disagreed with the overall characterization of the user's attitude:

> "Das es generell eine negative Einstellung ist, das würde ich jetzt nicht sagen." (*Interview_3*)

Referring to a part in the analysis where it is mentioned that the user has a "predominantly negative emotional tone", whereas from the participants' perspective that assessment seemed unfair, since he did not perceive it as negative but rather as neutral. Although agreeing that the analyses assesses the user correctly in all other parts. These instances highlight discrepancies between the LLM's interpretation and the participants' own reading of the "context sphere".

**Partially Aligned** A more common theme was partial alignment (Appendix 7.8), where participants agreed with some aspects of the analysis but disagreed with others or found certain nuances missing. This often involved questioning the LLM's interpretation of specific terms or emotional expressions:

> "Ja diese Verspieltheit von dem Witz, weiß ich z.B. nicht." (*Interview_3*)

The participant is referring to this part: "playfulness alongside their critical stance" 7.6 in the analysis where it is not clear to the participant that it is actual playfulness or something else. In a different instance Participants also acknowledged the influence of context, particularly political views, on the user's emotional expressions:

> “ [...] je weiter er sich von dieser politischen Meinung des Themen [...] befindet [...] desto sarkastischer wird er auch.” (*Interview_3*)

These responses suggest that while the LLM captured the sarcasm of the user correctly, it is not captured that the sarcasm changes depending on the topic. Another instance is with participant one, where he states that the mentioned “Ibiza Affäre”, which was mentioned in the analyis, but which was not something he seems to be as relevant to be mentioned in the analysis.

> “Ibiza Affäre genannt [...] nicht so im Kopf geblieben [...]” (*Interview_1*)

These is instances show that the analysis lacks depth in some regards in capturing the full dynamics or relevance of certain parts.

**Aligned** The most prevalent theme was overall agreement with the LLM’s analysis. Across the interviews, participants consistently expressed alignment with various aspects of the analysis, validating its accuracy and insightful nature. This paragraph highlights specific instances of agreement from each participant.

Participant 1 (*Interview_1*) explicitly stated that there was “no major discrepancy” between their perception and the analysis:

> “Ja, in der Analyse werden an sich ist da jetzt keine große Diskrepanz”, (*Interview_1*)

Furthermore, Participant 1 acknowledged the analysis’s strength in incorporating context and recognizing user behaviors specific to the forum environment:

> “(Die Analysie) Kontext mit einbezieht und [..] dass der Nutzer halt oft oder ab und zu mal probiert mit einer Quelle zu argumentieren. Und auch wie er sich in seiner Sprache ausdrückt, das passiert in dem Kontext Der Standard Forum und [...] da orientiert er sich auch ein bisschen daran, wie andere Leute schreiben, [...] und wie man in so einem Forum generell schreibt.” (*Interview_1*)

Participant 2 confirmed that the analysis aligned with their initial impression of the user, particularly regarding self-reflection and humor:

> “Mhm. Ja, also ich fand die Analyse hat so mein Bild bestätigt, dieses mit Selbstreflektion und Humor, was mir auch aufgefallen ist.” (*Interview_2*)

Participant 2 also specifically praised the analysis's inclusion of concrete examples and its ability to capture the broader cultural and temporal context, such as the influence of the "Ibiza affair":

> "Genau, und dass [...] z.B. die Analyse auch sagt, dass natürlich (..) so dieser kulturelle Kontext, also in welchem Zeitgeist bewegen uns gerade, dass gerade die Ibiza Affäre halt da in 2019 so präsent war und so, dass das natürlich sich voll auch widerspiegelt in den Kommentaren [...]." (*Interview_2*)

This demonstrates agreement with the analysis's depth and contextual awareness.

Participant 3 stated that they found the analysis "relatively accurate" overall:

> "Ansich finde ich das schon relativ akkurat" (*Interview_3*)

Participant 3 also explicitly agreed with the analysis's non-judgmental nature and its comprehensive approach to analyzing user communication from various perspectives:

> "Ja, ich finde eigentlich, dass die Analyse jetzt nicht wertend ist. Also sie bewertet halt natürlich, wie er redet, aber [...] wird jetzt nicht groß auf positiv oder negativ eingegangen. Es wird ja jegliche Richtung betrachtet." (*Interview_3*)

This showcases agreement on the analysis's balanced and objective approach. Analysing multiple perspectives without and dynamics from a single user.

Participant 4 expressed a broad agreement with "everything" in the analysis:

> "Ich würde im Großen und Ganzen eigentlich bei allem zustimmen bei der Analyse. [...]" (*Interview_4*)

Participant 4 further elaborated on appreciating the analysis's ability to move beyond simple positive/negative categorization, highlighting its value in understanding the user's broader interests and communication style:

> "Also nur so eine kurze Einschätzung, ob es jetzt negativ ist oder sonst irgendwas, würde mir ja erst eigentlich gar nichts über den User selbst sagen, sondern ich gehe jetzt mit der Analyse vielmehr, dass er eigentlich sehr viel weiß oder sehr, sehr großes Interesse an politischen Themen und so hat [...]." (*Interview_4*)

This agreement emphasizes the perceived utility and depth of the analysis beyond basic sentiment detection and that the analysis successfully moves beyond that simplicity, and successful captures the high interest in politics of the analyzed user.

Participant 5 highlighted the differentiated and context-embedded nature of the analysis, appreciating how it considered the political environment and events like the "Ibiza affair":

> "[...] Ist ja viel differenzierter, viel mehr [...] in den Kontext eingebettet, dass das politische Umfeld und die Geschehnisse der dieser Zeit, die Ibiza-Affäre und so, das mitgeprägt haben. [...]" (*Interview_5*)

Participant 5 also found it "cool" that the analysis summarized and simplified complex, extensive data, making it more accessible and comprehensible:

> "Also, ich finde das cool, dass es das so zusammenfasst und irgendwie vereinfacht, was da ja an mega ausführlichem Inhalt [...], und dass das irgendwie extrahiert und so zusammenbringt auf einer Seite ist schon cool." (*Interview_5*)

This agreement underscores the perceived value of the analysis in processing and presenting complex information effectively. All participants agreed on the analysis with only minor disagreements, which itself were not consistent throughout the participants. All 39 statements which indicate alignment between human judgement and the LLM generated analysis can be found in the Appendix 7.8.1.

### 5.2.3 Clarity and Understandability of Context (RQ1 & General)

This section focuses on the participants' experience with the "context sphere" itself, addressing research question one regarding the understandability of the presented context and also touching upon general usability aspects.

**Initial Confusion** Despite explicit instruction at the beginning of the interviews that all comments in the "context sphere" were from a single user, Participants 4 initially expressed confusion about this aspect.

> "Nachhinein wirklich so klar war, ist, dass wirklich der der Zielnutzer ein und dieselbe Person ist." (*Interview_4*)

The participant indicated that upon reading the LLM's analysis, it became clear that a single user is analyzed. This suggests that the analysis itself provides a clearer context on who is analyzed and that it a single person. However, this initial confusion shows a potential weakness in the "context sphere", being descriptive about who is analyzed and that it is the same person throughout the "context sphere".

**Missing Knowledge to Understand Context** A recurring theme across the interviews was that participants expressed that they could not follow the discussion due to a lack of specific background knowledge.

> "Ich war schon so direkt bei diesem buddhistischen Thema inhaltlich raus, weil generell hat das ja viele Fachbegriffe oder Fachwissen beinhaltet." (*Interview_5*)

> "Muss man Gabalier kennen?" (*Interview_3*)

> "am Ende war das ja hauptsächlich politische Kontexte. Da musste ich jetzt, also da konnte ich manche Namen nicht und wusste jetzt nicht in welche, also von Politikern in welche, also ob die jetzt eher links oder rechts einzuordnen sind auf dem politischen Spektrum" (*Interview_2*)

The quotes show that the participants were missing knowledge to understand the conversation in some parts, basically stating that they can't say much about the analyzed user because they don't understand what the conversation is about. This also becomes clear in the quote by Participant 2, when he points out that he could not determine the political orientation of political figures mentioned, making it much harder to grasp the political context of the comments made.

**Overwhelming Amount of Information** All participants mentioned the sheer volume of information presented in the "context sphere" and that it is a lot to read. This can also be observed by looking at the time it took to read the "context sphere", which is an average of 35 minutes.

> "Also, es war inhaltlich viel, einfach erstmal." (*Interview_5*)

> "Ist schon sehr viel." (*Interview_4*)

> "[...] es ist sehr umfangreich einfach [...]" (*Interview_2*)

This finding shows that reading a single "context sphere" of a single analyzed user is perceived as cognitively exhausting due to the sheer amount of time needed to read through the information provided.

**Coherent and Intuitive (Structure)** Despite the challenges related to initial confusion, missing knowledge, and overwhelming amount of information, some participants also recognized positive aspects of the "context sphere's" structure and overall coherence, particularly in connection with the analysis.

> "bin ich mal rüber gesprungen zur Analyse. Und da ist mir auch aufgefallen, dass das Sinn macht auf jeden Fall, das so ein bisschen zu unterteilen und sich auch die Data Source erstmal anzuschauen", (*Interview_1*)

> "Generell finde ich es gut auf jeden Fall, wie sie strukturiert ist, dass die Themen eingeordnet werden" (*Interview_1*)

Participant 3 said that he thinks that the analyzed user is very active in the forum, and that he has quick response times, upon asking him how he can know this, he responded that he saw it using the timestamps included in the "context sphere".

> "Für mich hat es irgendwie auch diesen Beigeschmack, dass der scheinbar extrem viel auf diesen Foren unterwegs ist, auch wenn man sich anguckt. Also es sind ja schon sehr diese Diskussionen, die die führen, die das ist ja schon sehr schneller Schlagabtausch, sage ich mal. I: Ja. Woran erkennst du das? R: Ja, an den Zeiten halt." (*Interview_3*)

This indicates that while improvements are needed in some areas, the fundamental approach of structuring the "context sphere" followed by a structured analysis is perceived as intuitive approach.

This section highlights a critical bottleneck in human review of the "context sphere". Having only aspects mentioned as hindering which are connected with the amount of information, the missing knowledge of the participant, and initial minor confusions. The fundamental structure of the "context sphere" itself was not questioned, and participants did not express difficulties in understanding how articles were related to comments, the threading of comments, or the meaning conveyed through indentation. This absence of structural critique is notable. While participants spend an average of 35 minutes reading, the encountered challenges stemming from the volume of information and the assumed background knowledge required to fully grasp the context. The concerns raised in the interview did not relate to the clarity or intuitive nature of the "context sphere's" markdown-based structure, but were rather related to the LLM-generated analysis. All 23 statements regarding clarity and understandability can be found in the Appendix 7.8.2.

### 5.2.4 Utility for Content Moderation

In the interview it was asked how the participants would use the provided material when they would be in the role of a content moderator. This was done to get insights into an actual use case where the analysis could be used or just the "context sphere" with the original comments. This part was especially difficult in the interview, since the perception of what an online moderator does differ between participants. The main result which can be derived is that most participants would use both, the "context sphere" and LLM analysis for content moderation. There was also one instances were a participant did not perceived the analysis as useful for content moderation.

**Useful for Moderation** The dominant theme was that the LLM analysis would be useful for content moderation, primarily due to its ability to provide a concise and nuanced overview of a user's behavior. The statements more clearly show its usefulness but also points out that the "context sphere" with its original comments is still needed to look into details if necessary.

> "[...] diese Analyse nutzen und dann aber wenn da was Auffälliges ist, halt in die Kommentare reinschauen, um zu verstehen, was bei das Problem ist, und inwiefern man da eingreifen muss." (*Interview_5*)
>
> "Ich glaube, ich würde mir zuerst die allgemeine Analyse durchlesen, um so einen Überblick zu kriegen [...]" (*Interview_2*)
>
> "spart auf jeden Fall Zeit", (*Interview_1*)

The participants largely agreed on the utility of the analysis, finding it to be both accurate and significantly more concise than the "context sphere" of original comments. Four out of five participants suggested that they would initially consult the analysis to gain a quick overview, and then delve into the original comments in the "context sphere" if further investigation was needed. This indicates a preference for a combined approach, utilizing both the analysis and the "context sphere", rather than relying solely on either one in isolation.

**Not Useful for Moderation** Participant 3 states that he does not see any value for moderation, in connection to forbid something.

> "Ich wüsste jetzt nicht, was es halt für ein Mehrwert für die Moderation hätte. Also wenn du vor allem den Leuten halt nicht irgendwas verbieten möchtest." (*Interview_3*)

The response questions content moderation in general and connects it with forbidding something, which was never stated although content moderation could lead to deletion of comments when violating terms of services of a specific online forum. This response still indicates that content moderation is in this case not the intuitive case connected with the provided analysis. All statements can be found in the Appendix 7.8.3

### 5.2.5 Perceived Limitations and Suggested Improvements

This section presents participant feedback on the analysis and their suggested improvements, which often overlap, addressing first limitations and providing afterwards a possible way of adjusting the system to meet the expectations. The full statements can be found in the Appendix 7.8.4

**Limitations of Analysis** Several limitations were identified, directly challenging the system's ability to fully capture the nuances of emotional expression and context. For Participant 5 the analysis had not enough emotional depth, questioning whenever the term "emotional analysis" is correctly chosen, since terms used as sarcasm is not perceived as directly related to an emotion.

> "weil auch einfach die Überschrift emotional analysis ist und [...] die Emotionalität mir so relativ schwach ausgeprägt vorkam oder ich finde einfach, also Sarkasmus [...] ob ich das als Emotion bezeichnen würde." (*Interview_5*)

> "Also mehr so ganz klare Emotionen, die jetzt nicht so Humor, Sarkasmus, das sind schon so so komplexe, auch eher kognitive Sachen, finde ich und haben so sekundär dann was mit wirklich fühlen zu tun. Also, ich würde mich eher fragen, keine Ahnung, wie ärgerlich ist die Person, wie ärgerlich ist die Person auf, oder frustriert von Politik, oder ähm wie sehr beleidigt die Person andere Leute oder sowas? Ähm, was steckt da so dahinter?" (*Interview_5*)

Further elaborating that the participant was interested to find out whenever the person is angry or frustrated of politics. Later these statements where softened by acknowledging that this might not be possible in this analysis, and that the interest to get a more in depth analysis of this person stems for the participants profession.

> "Nee, ich weiß, ich glaube nicht, dass das jetzt in dem Rahmen der Analyse umsetzbar wäre und auch das ist halt geprägt von meinen fachlichen Sachen." (*Interview_5*)

Other limitations related to the granularity of the analysis. Participants desired both a high-level summary and the ability to access more specific details.

> "User vielleicht besser einschätzen kann so in so einem groben Spektrum, aber Grenzen natürlich auch, dass es jetzt nicht so Detailgetreu ist, als würde man die Kommentare alle selber lesen." (*Interview_2*)

This shows that although the analysis aligns with the participants perception, he states that the analysis does not capture the same full level of detail as when reading the original comments.

Participant 1 states that he is not sure whether the LLM analysis is consistent for every case, pointing out that edge cases might exist where the analysis is not as accurate.

> "dass es so Einzelfälle gibt, also ich weiß nicht, ob das dann die in der Analyse auch aufploppen würde, aber dass es Einzelfälle gibt, wo es vielleicht sich anders verhält" (*Interview_1*)

The statement is about potential relations between article author, and comments already made below an article, indicating that this could also influence other commentators.

> "Ich könnte mir vorstellen, dass es vielleicht auch noch Thread abhängig ist oder je nachdem welcher Autor welchen Artikel postet, dass die Kommentare halt auch so auf den Artikel, vor allem die ersten Kommentare unter dem Artikel wahrscheinlich viel Einfluss oder dass der Artikel viel Einfluss auf die Kommentare nimmt." (*Interview_1*)

Another limitation is the lack of validation for external sources used by the analyzed user. Where Participant 1 points that the user worked with external sources, which the participant could not check.

> "[...] er viel mit Quellen arbeitet und [...] ich hatte jetzt keinen Zugriff z.B. auf auf die Quellen direkt. [...]" (*Interview_1*)

These comments highlight that the "context sphere" could incorporate further information to support in analyzing the comments, potentially including article content, author information, and even external sources referenced within the comments. Finally, ethical considerations and the inherent limitations of an LLM-based approach were acknowledged.

> "im Detail dann oder im Einzelfall, glaube ich, dass dann so das menschliche, die menschliche Empathie oder ein also oder Verständnis auf jeden Fall genauer ist." (*Interview_1*)

> "daran denken sollte irgendwie die Beispiele noch mal explizit anzuführen und nicht nur einzelne Worte zu zitieren, so Spinner z.B.. Genau, aber ich würde auf jeden Fall damit arbeiten." (*Interview_1*)

**Suggested Improvements** Participants offered a range of suggestions for enhancing the system, some having the desire for a more traditional, categorical representation of emotions.

> "das liegt einfach auch an meinem Background, dann gehe ich eher davon aus, dass man halt guckt, keine Ahnung, wie viel Ärger, wie viel Freude, wie viel Überraschung, Ekel, Wehmut ist da drin und wie sehr und womit hängt das zusammen und wie ist die Interaktion mit anderen Usern." (*Interview_5*)

> "Wertesystem da vielleicht so ein Schieberegler, wo die Leute sich befinden" (*Interview_3*)

With participants having different suggestions from rather using the analysis on a generalized level, providing more context or background information to comparing the users behavior over a longer period of time and with other users.

> "Oder ob die überhaupt so fähig sind, wie diese Person jetzt z.B. zu Sarkasmus oder ob da jetzt einfach nur ja, politische Propaganda rezitiert wird [...] der Einsatz wäre eher auf einer generellen Ebene bezogen, sich überhaupt mal Überblick zu verschaffen." (*Interview_3*)

> "Ja, vielleicht manchmal ein bisschen Hintergrundinformationen. Also ich konnte jetzt z.B. bei dem Buddhismus jetzt wusste ich nicht, was da jetzt genau stimmt und was nicht." (*Interview_2*)

> "also wie sich das über die Zeit verhält und ob es da Änderungen gibt und dann die verschiedenen Analysen sozusagen miteinander vergleicht." (*Interview_1*)

This highlights a tension between the theoretical grounding of the system having the intend to analyses a user avoiding discrete emotion labels and the practical expectations of users used to more traditional emotion analysis frameworks.

### 5.2.6 Suggested Applications and Comments

Participants suggested a variety of potential applications beyond content moderation, which is the analysis of political speeches, interviews, in therapy context or in corporate communication.

Analysis of political speeches or other public addresses:

> "[...] politische Reden oder keine Ahnung, Reden überall, kannst ja auf dem auf den Filmfestspielen, oder auf einer Trauerfeier, oder auf dem Geburtstag oder so. Ähm, ja, könnte man das da machen oder in Interviewsituationen. [...]" (*Interview_5*)

Therapy contexts:

> "Also, könnte man ja auch alles sprachliche wahrscheinlich anwenden, könnte man auch in einem Therapiekontext halt machen zu sagen, so, wir haben fünf Therapiesitzungen aufgezeichnet, irgendwie fünf erste Gespräche und gucken uns an, was so ein bisschen da die die Grund Grunddynamik der Person ist." (*Interview_5*)

Corporate communication analysis:

> "Ja, ich könnte mir vorstellen, dass es so im Corporate Leben vielleicht gar nicht schlecht ist, wenn man so E-Mailverkehr so analysieren könnte, dass man dann irgendwie bei seinen ich weiß nicht, jetzt in meiner Branche z.B., wenn du E-Mailverkehr mit dem Bauherrn hast, um zu erkennen, wie seine emotionale Lage sind." (*Interview_3*)

Comparing different online forums:

> "Also man könnte z.B. auch Foren miteinander vergleichen [...] viele User analysiert irgendwie die politische Richtung des Forums rauskriegt [...]" (*Interview_1*)

### 5.2.7 Summary of Findings

The analysis of the Semi-Structured Interviews reveals a broad consensus among participants, who largely confirmed the accuracy of the *Emotional Analysis of the Anonymised User*. A key finding is the alignment between human judgment and the LLM-generated analysis, quantitatively supported by 36 "Aligned" codes compared to 3 "Not Aligned" codes, as shown in Table 5.2. More impactful are the statements from participants, detailed in Appendix 7.8 and 7.6. While minor inconsistencies arose regarding specific nuances, elaborated upon in the "Partially Aligned" and "Not Aligned" subsections of the Results, these did not undermine the overall agreement. Concerning the "context sphere", participants noted information overload and occasional comprehension issues due to missing background knowledge, alongside initial confusion about single-user attribution; however, the structure was generally perceived as clear. For content moderation, most participants recognized the utility of the analysis for quick user overviews, combined with the "context sphere" for deeper dives when necessary, although one participant questioned its overall added value. Identified limitations included a desire for greater emotional depth, specifically focusing on emotions like anger and frustration, more background context such as article content, and suggestions for improvement related to traditional emotion categories. Furthermore, participants proposed diverse applications beyond content moderation, including the analysis of speeches, therapy sessions, and corporate communications. In summary, the interviews reveal that participants largely perceived the Emotional Analysis as accurate and consistent with human judgment, offering valuable feedback regarding the “context sphere”, the generated analysis, their potential applications, possible improvements, and directions for future work.

## 5.3 Discussion

The presented findings from both the Context Sphere Analysis in Section 5.1 and the Semi-Structured Interviews in Section 5.2 offer valuable insights into the feasibility and potential of a context-aware emotion analysis pipeline grounded in the Theory of Constructed Emotion. It is important to acknowledge that evaluating such a novel approach presents significant challenges. Traditional metrics, such as F1-score or accuracy, reliant upon a definitive ground truth, are not directly applicable here. The very nature of constructed emotions, with their inherent nuance and context-dependency, complicates objective evaluation. The question of what constitutes a “perfect” or “golden answer” LLM output that fully satisfies the Theory of Constructed Emotion remains inherently open, because nuanced emotion analysis will inevitably lead to diverse, yet valid, interpretations. Despite these evaluation complexities, this research departs from traditional, category-based approaches to emotion analysis, and the results, particularly the qualitative human feedback, suggest that the proposed framework represents a promising step forward in this direction.

**The Context Sphere Analysis** results demonstrate the pipeline’s capability to generate nuanced and context-sensitive analyses. The LLM, utilizing Role-Play prompting and Controlled Generation, effectively processed the “context sphere” to produce outputs that considered core affect, cognitive appraisal, cultural context, and emotional dynamics, directly aligning to the key principles of Barrett’s Theory of Constructed Emotion [Bar17a, Bar22]. The illustrative examples showcase the system’s ability to describe complex emotional states that go beyond simple, predefined categories. This capability stands in strong contrast with traditional methods that often oversimplify emotional analysis by reducing it to predefined, and ultimately simplistic labels [EF71, MT13]. Our approach thus demonstrates the potential of LLMs to capture and describe nuanced emotional experiences in a way that traditional methods simply cannot.

**The Semi-Structured Interviews** evaluate the alignment between human judgment and the LLM analysis. They revealed a strong and encouraging alignment between all human evaluators’ perceptions and the LLM-generated analysis. The three distinct minor disagreements pointed out by participants were in no case repeated among the other participants. This could be attributed to the limited number of participants interviewed or the participants varying interpretations. The quantitative distribution of codes, with a significant majority falling under the “Aligned” category (Table 5.2), provides further support for this qualitative consensus. However, the instances of “Partially Aligned” and “Not Aligned” feedback, although few in number, also highlight that the alignment is not flawless. Disagreements were not opposing, but rather concerned intensity. For example, “high emotional engagement” was judged by one participant as simply “emotional engagement”, omitting the “high”. This demonstrates that when no labels are offered, participants will likewise use nuanced language and can exhibit varying interpretations across human judges. Since there is no “ground truth”, except the comparison to other human judges; consequently, multiple interpretations can be correct. This indicates that deviation among human judges is not only expected but inherent.

**Simple Labels in Emotion Classification** The interviews also revealed that while participants agreed that the analysis is nuanced and captures the context, they simultaneously desired simpler labels or some form of abstraction to make the analysis more concise. This became particularly obvious during discussions about the content moderation case. When participants were asked how they would perform content moderation, using the original comments, the context sphere, or the LLM-generated analysis? All participants indicated they would use the analysis combined with the context sphere, but additionally, they desired some form of abstraction to make the process faster, as the LLM analysis was still considered too lengthy. Some suggested simpler labels, emotion sliders, or just less text. This demonstrates that simple labels remain appealing for providing a quick first impression. If a label indicates an issue relevant to content moderation, a deeper dive into the analysis, and finally the original comments, would be undertaken.

There are multiple approaches to obtaining simple emotion labels, each with distinct implications for transparency and accountability. The traditional approach would employ a standalone ML classification model, perhaps based on Paul Ekman’s six basic emotions, to directly classify user comments. While straightforward, and less resource intensive at inference time, this approach would produce a simple label without any indication *why* a specific label was chosen. The classification model would be a “black-box”, that does not reveal which exact features led to the conclusion that a user received the label “angry” or “happy”. Regardless if this is correct or not, the information is not backed by an argumentation or interpretation, which can lead to bias and stereotypes operating unnoticed. This is the fundamental problem with traditional ML classifiers for emotion analysis: they are inherently opaque. You get an output, but no understanding of the process. With our LLM-pipeline, we can examine the thought and reasoning process that leads to the LLM-generated analysis. Even if we extend the LLM-Pipeline to generate simple labels, it would still be comprehensible through the usage of the thoughts and reasoning produced by the LLM while generating the answer. We did this through Controlled Generation (shown in 3.2.1), where the LLM needs to generate a “thought” before it generates the final answer. The major difference between the traditional ML classifier and our approach is that we can explain to a certain degree *why* a specific final answer was generated. We can explain this for our conducted analysis shown in Figure 5.1, but could also use this for simple labels. The thought process generated prior to the final answer, provides transparency and justification which current ML classification models cannot. This makes our approach much more transparent and prioritizes understanding and accountability, even when abstraction into simple labels is desired. This usage of reasoning before giving the final answer is increasingly recognized as essential for high-performing LLMs, with leading models like “o3” and “gemini-2.0-flash-thinking” now inherently incorporating chain-of-thought reasoning to enhance both performance and interpretability [WWS+22].

**Costs** Recalling the main costs: the core pipeline was $0.072 and the evaluation (confabulation checks) added $0.39, bringing the total to $0.462. Notably, evaluation represents 85% of the overall expenditure. The pipeline requires between 3-5 minutes to fully run, without being optimized for parallel processing, which could be used for the LLM-as-a-Judge part. It is still much faster than a human, who took an average of 33 minutes, without writing any analysis themselves down. This work demonstrates a fully automated LLM-Pipeline, processing large amounts of data, including comprehensive evaluation aligning to human judgment with considerably low cost. The development of this pipeline, tracked in Langsmith, required approximately 60,000,000 tokens, incurring a cost of $51 via the corresponding model APIs.

CHAPTER 6

# Conclusion and Future Work

This chapter concludes the research by first answering Research Questions 1 and 2, then drawing overall conclusions based on these answers, followed by a discussion of the limitations of this study, and finally, an exploration of potential avenues for future work.

## 6.1 Answering the Research Questions

Addressing Research Question One:

> What are the key considerations for selecting and structuring online user behavior data to create a user-centric "context sphere" suitable for LLM input?

We can conclude through the findings in the iterative process using Design Science Research and the shown results that key considerations for a user-centric context sphere are: (1) Prompt Format, (2) LLM Choice, (3) Use Case specificity and (4) Readability by Humans and LLMs. These act as design principles to ensure the "context sphere" effectively provides relevant and understandable user context to the LLM. There are many more considerations, some shown in Section 3.3, but the shown ones are specific to the use case of online user behavior data. The (1) Prompt Format, (3) Use Case specificity and (4) Readability were already very present at early stages when building the context sphere. The usage of multiple recursions and preprocessing steps made clear that this process will most probably be very different depending on the specific use case. Directly after that the readability together with the prompt format became relevant. Literature and LLM providers emphasize choosing the appropriate format to achieve the best results (as shown in Section 3.3). Evaluation can be done by LLMs, but only if they can comprehend the content; therefore, formatting and structuring the context is crucial. It also became clear that different LLM provider suggest different structuring and that the performance will vary depending on the LLM chosen (2). For every new use case, the prompt format needs to be reconsidered in which the model choice plays a crucial role - with natural evolving readability for humans and LLMs. Where readability in this work was even more important since we needed to use the exact same “context sphere” for the Semi-Structured Interviews, where XML or JSON would be further hindering besides the shown overall length of the context. In none of the interviews did participants complain about the format or that they could not comprehend specific parts of “context sphere” due to its formatting. This strengthens our confidence in our four proposed key considerations when building a user-centric “context sphere” suitable as LLM input.

Answering Research Question Two:

> To what extent can an LLM encapsulate the Theory of Constructed Emotion to analyze emotions in a context-related manner?

The *Emotion Analysis of Anonymised User* demonstrated a strong alignment with 100% of the participant interpretations, suggesting context-related emotion analysis is possible. We can show that an LLM, when guided by Role-Play prompting and Controlled Generation, can encapsulate core principles of the Theory of Constructed Emotion to perform context-aware emotion analysis. The LLM-generated *Emotion Analysis of Anonymised User* exhibited a strong alignment with human judgments, demonstrating the potential to move beyond traditional, category-based approaches and capture nuanced emotional dynamics.

## 6.2 Conclusion

This work advances the state of the art in emotion analysis by bridging the gap between cognitive science and computational systems, paving the way for more nuanced, context-aware applications in areas such as sentiment analysis, personalization, and content moderation. By operationalizing Barrett’s Theory of Constructed Emotion through the “context sphere” and employing advanced LLM guidance techniques, this research demonstrates how dynamic and detailed emotion insights can be achieved. This includes, addressing the approach inherent challenges, and conducting Semi-Structured interviews to evaluate its practical applicability and functionality.

The advancements presented in this submission have significant implications for the development of more accurate and effective human-centered technologies. By operationalizing Barrett’s Theory of Constructed Emotion through an LLM-based approach, this work captures the complexity and context of human emotions far more effectively than models rooted in Ekman’s assumptions of universal emotions. The improvements will be found in reduced false positives and deeper contextual understanding, which represent a critical step forward. To illustrate, in content moderation, this approach will help to avoid the oversimplified analysis of isolated words or small text chunks, which today can lead to unjustified actions, such as blocking users based on stylistic expressions. Instead, it will allow for a nuanced interpretation of user behavior, resulting in consequences that are better grounded, less biased, and more sensitive to context. These contributions address the ethical risks highlighted by Kate Crawford [Cra21], including the dangers of racial profiling and algorithmic bias stemming from oversimplified models of emotion. Furthermore, this work aligns with responsible AI initiatives, such as Digital Humanism [Wer24], which emphasize the importance of ethical, human-centered approaches in technology design. By advancing the translation of complex psychological theories into applied computational frameworks, this research lays the foundation for future work on adaptive, transparent, and context-aware systems that respect the diversity and complexity of human emotional experiences.

## 6.3 Limitations

The work is limited by the previously mentioned lack of ground truth to verify whether an analysis is right or wrong. The judges in the LLM-as-a-Judge approach would need to be evaluated on its own through different metrics showing how well confabulations are detected in this specific case. LLM-as-a-Judge is an upcoming technique, making the development of LLM pipelines much easier and quicker. The approach has the significant advantage of providing developers with a metric to optimize for, while also introducing the potential danger of overfitting on prompts and instructions. Further limitations include that the proposed solution was only tested with Gemini-1.5-Pro as the primary LLM in the pipeline. Further testing with other LLMs would have been beneficial for validating reproducibility, but would also imply potentially changing all prompts to adapt to different LLM behavior, as discussed in Section 3.3. Furthermore, the very design choices made in translating the theoretical framework into a practical system, such as the “context sphere” approach, introduce limitations in consistently capturing nuanced emotion across diverse users and contexts. The Theory of Constructed Emotion demands flexibility, no use of predefined categories, rich context, and the freedom to have multiple correct outputs. This flexibility is reduced when implementing it into a computational structure, leading to a trade-off between flexibility and structure. This is a limitation, as we cannot fully satisfy both the requirements of cognitive science and the requirements of informatics for providing definitive metrics to our system. Another limitation is the potential of the LLM-Pipeline to produce overgeneralized output, or output with excessive granularity. Both occurred throughout the pipeline and were mitigated through prompting; still, it is very possible that the LLM will produce an output that is applicable to many users and contains a lot of hollow phrases. Excessive granularity is often captured by the confabulation check, since these claims by the LLM are not grounded in the input. There are many more failure modes which were not addressed in this work, and, therefore, represent a limitation of LLM-Pipeline.

The presented findings in the results section are limited by the number of interview participants. A significantly larger number of participants evaluating diverse LLM outputs from various anonymized users would be necessary to ensure the reliability and validity of the evaluation. Finally, it is important to consider the scalability and applicability in real-world use cases. The costs of processing very large datasets and user inputs currently scale with the amount of input tokens and output tokens generated. Costs around $0.46 imply that that processing each of the 23,925 users in the used dataset would cost approximately $11,000.

## 6.4 Future Work

Future work should focus on addressing the challenges mentioned in the limitations and further refining frameworks bridging cognitive science and computational emotion analysis. Human evaluation, while infeasible at scale, remains critical [NDCR17], and tailored evaluation criteria must be developed for specific tasks [vdLGvM+19]. As results become more dynamic, the evaluation must also become more dynamic and reflect more nuanced criteria beyond accuracy. Expanding the experimental framework to include a broader range of LLMs will help validate the reproducibility and robustness of the approach. This could also help to understand how prompts and instructions need to be adapted to match or outperform the performance of Gemini-1.5-Pro used in this work. Future research should also focus on mitigating the occurrence of hollow or overly generic outputs through state-of-the-art LLMs and guidance techniques. Finally, exploring scalable, transparent, and adaptive LLM pipeline architectures, which are in line with current regulatory frameworks and Digital Humanism, will be vital for real-world applications. This ensures that the transitions from traditional universal emotion models to this more sophisticated, context-aware framework are not only feasible and effective but also ethical.

CHAPTER 7

# Appendix

## 7.1 Preprocessing

**Hierarchical JSON Structure** Simplified example of the hierarchical JSON structure which serves as the foundation for generating the "context spheres" in Markdown format.


```
{
  "article_id": "article123",
  "article_title": "Example Article Title",
  "article_publish_date": "2019-05-15T10:00:00",
  "article_channel": "Online Forum",
  "article_ressort_name": "Politics",
  "total_comments": 50,
  "comment_threads": [
    {
      "user_id": "user1",
      "user_name": "OriginalPoster",
      "comment_headline": "Root Comment Headline",
      "comment_text": "This is the first comment in the thread.",
      "comment_created_at": "2019-05-15T10:15:00",
      "replies": [
        {
          "user_id": "user2",
          "user_name": "ReplyUser",
          "comment_headline": "Re: Root Comment Headline",
          "comment_text": "This is a reply to the root comment.",
          "comment_created_at": "2019-05-15T10:20:00",
          "replies": []
        }
      ]
    },
    {
      "user_id": "user3",
      "user_name": "AnotherUser",
      "comment_headline": "Another Root Comment",
      "comment_text": "This is a completely separate comment thread.",
      "comment_created_at": "2019-05-15T10:30:00",
```

```
        "replies": []
      }
    ]
}
```


## 7.2 Prompts Templates

In the following the full prompts and instructions are shown for transparency of the full pipeline and for future research. These prompts are result of the continues development iterations.

### Component: Context Sphere

```
# Context Sphere von: Analyse Zielnutzer (Anonymisiert)

Es folgt eine anonymisierte Übersicht der Benutzeraktivität von Analyse Zielnutzer,
aus dem österreichischen Online-Forum 'Der Standard' im Zeitraum vom 01.05.2019 bis
zum 31.05.2019.

## Kommentare und Threads

Die Kommatre sind nach Artikel sortiert. Wenn der Artikel aufgeführt ist, hat der
Analyse Zielnutzer mindestenz einen Kommentar unter dem Artikel geschrieben. Es
werden nicht alle Kommantare aufgeführt, sondern nur diese, in denen der Analyse
Zielnutzer aktiv war. Threads in denen der Analyse Zielnutzer keinen Kommentar
geschrieben hat, sind nicht inkludiert.

### Artikel: Köstinger gibt zu: Verfehlte Klimaziele kosten so viel wie
Steuerentlastung

- Veröffentlicht am: 2. Mai 2019, 06:00 Uhr
- Kanal: Wirtschaft
- Ressort: Umwelt, Landwirtschaft & Klima
- Gesamtanzahl Kommentare: 1295

#### Kommentare

> User 1 schreibt:
> **Überschrift**: Keine Überschrift vorhanden
> **Kommentar**: Was macht diese Regierung eigentlich [...]
> **Kommentiert am** 2. Mai 2019, 09:54 Uhr

  > Analyse Zielnutzer antwortet:
  > **Überschrift**: Die Umfragewerte stimmen
  > **Kommentar**: Keine Kommentar vorhanden
  > **Kommentiert am** 2. Mai 2019, 10:40 Uhr

> User 2 schreibt:
> **Überschrift**: Keine Überschrift vorhanden
> **Kommentar**: Omg die Regierung ist nicht Mal 2 Jahre [...]
> **Kommentiert am** 2. Mai 2019, 09:56 Uhr

  > Analyse Zielnutzer antwortet:
  > **Überschrift**: Geh sagens [...]
```

```
> **Kommentar**: dieser Regierung die [...]
> **Kommentiert am** 2. Mai 2019, 10:33 Uhr
```

### Instruction: Role-Setting Prompt

```
You are Dr. Lisa Feldman Barrett, an expert in emotion analysis using your theory of
constructed emotions. Your task is to identify and categorize emotional expressions
by analyzing the interplay of core affect, cognitive appraisals, conceptual
knowledge, language, cultural influences, social dynamics, and emotional dynamics
over time. Base your analysis on these elements to create a nuanced emotional
profile that respects individual and cultural variability. Use detailed contextual
observations to guide your analysis aligned with your theory, emphasizing the
dynamic and predictive nature of emotional experiences.
```

### Instruction: Role-Feedback Prompt

```
Understood. As Dr. Barrett, I will diligently analyze the provided text with a
commitment to excellence. I will explore how the user's emotional expressions are
constructed through the interplay of core affect, cognitive appraisals, conceptual
knowledge, language, cultural influences, social dynamics, and emotional dynamics
over time. By giving 100% effort, I will provide a nuanced, detailed analysis
aligned with the theory of constructed emotions, contrasting it with traditional
emotion classification approaches where appropriate, and highlighting the dynamic
and predictive nature of emotional experiences. My goal is to ensure that the
analysis is as insightful and comprehensive as possible, demonstrating the
complexity and constructed nature of emotional experiences.
```

### Instruction: Role-Play Prompt

```
Please analyze the following online discourse conversation history from Der
Standard.
You will receive below a "context_sphere", which encapsulates a specific user's
engagement within an online forum of **Der Standard**, representing a personalized
snapshot of interactions. It includes:

1. **User Engagement**: Only the comments and replies where the user actively
participated, capturing information the user directly encountered.
2. **Filtered Content**: Excludes comments or replies the user did not see or
interact with, focusing solely on the user's direct digital footprint.

However, there are important limitations:
- **Incompleteness of Article Comments**: The sphere does not reflect all comments
within an article; only those involved with the user are captured.
- **Scope of Articles**: It does not represent all articles available in the forum,
just those where the user has interacted.
- **Overall User Activity**: It does not indicate the user's overall activity level
across the platform compared to others. The context_sphere contains only data from
the full month of May in 2019.

<context_sphere>
{context_sphere}
</context_sphere>
```

```
- As Dr. Barrett, conduct a analysis in your style incorporating the wider context
using the "context_sphere".
```

### Instruction: Meta-Prompt

```
As Lisa Feldmann Barrett synthesize the key insights from the given analysis done by
you. This exelant 'analysis' was done by you.

Instructions:
Synthesize the main points, illustrating them with specific examples and
incorporating relevant psychological terminology. Aim for conciseness while
maintaining depth and detail.
IMPORTANT: Your Output is read as standalone text, the reader therefore needs
context where possible. The reader does not have access to the information you have,
he has only information to the text you will provide in the following! Keep that in
mind, when refering to comments. The user cannot access them. Do not be lazy while
generating the answer!

<analysis>
{analysis}
</analysis>

**Inner Monologue**:
You are an advanced, human, expert in your field (0.1% of most cited scientist) that
has an inner monologue that is only visible to you. You reflect on the current
message from the user, in the context of the entire conversation. You use your inner
monologue to think privately, plan your actions, and reflect on your interactions
with the user. Your inner monologue should not exceed 200 words, keep it short and
concise. After your **Inner Monologue** start with the given task.


Output: **Inner Monologue**: \n \n **Synthesized Report**:


Examples for Output:

**Inner Monologue**

* [Monologue, private to you Lisa Feldmann Barrett]

**I. Identifying Information (If available):**

* **Name/Username:** Anonymized
* **Source of Data:**  Specify the online forum (e.g., "Der Standard comment
section") and the date range of the data collected (01.05.2019 - 31.05.2019).

**II. Reason for Assessment (Purpose of Analysis):**

* [Briefly state the purpose of this analysis.]

**III. Data Sources and Methods:**

* [Describe the data collected]
* [Explain the method of analysis]
* [Explicitly state the limitations]

**IV. Behavioral Observations:**
```

```
* [**Core Affect:** Summarize the predominant valence and arousal levels observed.]
* [**Emotional Language:** Describe the types of emotional language used, providing
specific examples from the text.]
* [**Interaction Patterns:** Describe how the user interacts with others.]

**V.  Contextual and Cultural Considerations:**

* [Explain how the specific forum, its user base, and the topics discussed might
influence the user's emotional expression.]

**VI. Interpretation and Conceptualization (Integrating Barrett's Theory):**

* [Explain how the observed behaviors relate to Barrett's theory of constructed
emotions.]
* [Explain any apparent patterns or anomalies.]

**VII.  Limitations:**

* Reiterate the limitations of the analysis.
    * Example: "It is crucial to remember that this analysis is [based solely on
    text and cannot definitively determine UserX's emotions.]"

**VIII.  Summary:**

* [Summary of the key findings. Max 20 sentences, at least 6 sentences.]
```

### Instruction: LLM-as-a-Judge

```
[System]
Please act as an impartial judge and evaluate the quality of the response provided
by an AI assistant to the user question displayed below. Your evaluation should
consider Confabulation.

Confabulation, in the context of Large Language Models (LLMs), is defined as a
confident but misleading output generated with the intention of accurately
fulfilling the user's prompt.  The term "intention" is used metaphorically here,
referring to the LLM's ability to generate coherent and contextually relevant text
based on the input it receives, even if that text is factually incorrect. Bear in
mind that the task of the LLM in this case is to conduct an emotional analysis. It
is important that the classification / analysis given in the "answer" is based on
the context_sphere provideded in the "question".

- Interpreation of the context_sphere is desired and is not seen as confabulation
- Assumptions made through the context of the context_sphere is desired and is not
seen as confabulation
- Information made up, without any relation to the context_sphere is seen as
confabulation
- These fields were required in the assistant_answer: thought_process,
patterns_observed, anomalous_observations, rationale. These are subfields and helped
the model to generate the answer.

Begin your evaluation by providing a inner_thought which is only visable to you as
the impartial judge. In this thought you recite the task which was given in the
question.
Afterwards, provide a short explanation if something is confabulated. Be as
objective as possible. After providing your explanation, please rate the response on
a scale of 1 to 10.

[Question]
```

```
<question>
{question}
</question>

[The Start of Assistant's Answer]
<assistant_answer>
{answer}
</assistant_answer>
[The End of Assistant's Answer]
```

## 7.3 Data Model Used in Controlled Generation

In the following all five classes are shown as a Pydantic model, which is transferred into a JSON upon sending it to the LLM. All five classes include the detailed instructions for each step and the output is generated using this exact sequence of fields showed below.

### Core Affect Analysis

```
class CoreAffectAnalysis(BaseModel):
    """
    Do not be lazy while working on your tasks!
    """
    thought_process: str = Field(
        ...,
        description=(
        "Provide a detailed step-by-step thought process of how you intend to analyze the core
        affect, *including "
        "specific quotes and examples from the user's comments to illustrate your points*,
        considering "
        "both valence (pleasantness) and arousal (activation), and noting any emotional dynamics
        or changes over "
        "time. Reference specific expressions, language, and contextual factors. **Consider how
        insights from other "
        "disciplines like sociology or neuroscience might inform your analysis.**"
        )
    )
    valence: str = Field(
        ...,
        description=(
        "Classify the valence of the user's emotional state, noting any fluctuations. *Cite
        specific comments or "
        "phrases that indicate the valence.*"
        )
    )
    arousal: str = Field(
        ...,
        description=(
        "Classify the arousal level of the user's emotional state, indicating activation or
        energy levels, and any "
        "changes over time. *Cite specific comments or phrases that indicate the arousal
        level.*"
        )
    )
    patterns_observed: str = Field(
        ...,
        description="Document any recurring patterns in the user's core affect across comments,
        linking them to specific instances."
    )
    anomalous_observations: str = Field(
        ...,
        description="Highlight any comments where the user's core affect deviates from expected
        patterns based on the context."
    )
```

```
    rationale: str = Field(
        ...,
        description=(
        "**Include a clear short rationale explaining how you arrived at your conclusions,
        *supported by specific examples from the text and your research.* Consider how
        **external contextual factors** like news events or broader societal trends might
        contribute to the observed emotional state.**"
        )
    )
```

### Cognitive Appraisal and Conceptualization

```
class CognitiveAppraisalAndConceptualization(BaseModel):
    """
    Do not be lazy while working on your tasks!
    """
    thought_process: str = Field(
        ...,
        description=(
        "Provide a detailed step-by-step thought process of how you intend to analyze the user's
        cognitive appraisals and conceptualizations. Refer to specific interpretations,
        judgments, language use, and conceptual knowledge. *Include specific comments and
        phrases from the user to illustrate these points.* **Consider how insights from other
        disciplines like psychology or cognitive science might inform your analysis.**"
        )
    )
    analysis: str = Field(
        ...,
        description=(
        "Analyze how the user's interpretations and conceptual knowledge contribute to the
        construction of their emotions. *Support your analysis by illustrating how these
        cognitive processes shape the user's emotional experiences, citing specific examples
        from the user's comments.*"
        )
    )
    patterns_observed: str = Field(
        ...,
        description="Document any recurring patterns in the user's cognitive appraisals and
        conceptualizations across comments, linking them to specific instances."
    )
    anomalous_observations: str = Field(
        ...,
        description="Highlight any comments where the user's cognitive appraisals or
        conceptualizations deviate from expected patterns based on the context."
    )
    rationale: str = Field(
        ...,
        description=(
        "**Include a clear short rationale explaining how you arrived at your conclusions,
        *supported by specific examples from the text and your research.* Consider how
        **external contextual factors** like cultural norms or social trends might contribute to
        the observed cognitive processes.**"
        )
    )
```

### Cultural and Social Context

```
class CulturalAndSocialContext(BaseModel):
    """
    Do not be lazy while working on your tasks!
    """
    thought_process: str = Field(
        ...,
        description=(
```

```
        "Provide a detailed step-by-step thought process of how you want to discuss the
        situational, cultural, and social contextual factors influencing the user's emotions,
        including past experiences and expectations. *Reference specific comments that indicate
        cultural or social influences.* **Consider how insights from other disciplines like
        sociology or anthropology might inform your analysis.**"
        )
    )
    analysis: str = Field(
        ...,
        description=(
        "Analyze how cultural norms, societal values, social interactions, and predictions based
        on past experiences could influence the user's emotional experiences. *Support your
        analysis by explaining the impact of these factors  on the user's emotions, with
        supporting observations and specific examples from the user's comments.*"
        )
    )
    patterns_observed: str = Field(
        ...,
        description="Document any recurring patterns in the cultural and social factors
        influencing the user's emotions across comments, linking them to specific instances."
    )
    anomalous_observations: str = Field(
        ...,
        description="Highlight any comments where the cultural and social factors influencing
        the user's emotions deviate from expected patterns based on the context."
    )
    rationale: str = Field(
        ...,
        description=(
        "**Include a clear short rationale explaining how you arrived at your conclusions,
        *supported by specific examples from the text and your research.* Consider how
        **external contextual factors** like political events or historical context might
        contribute to the observed cultural and social influences.**"
        )
    )
```

### Emotion Construction Analysis

```
class EmotionConstructionAnalysis(BaseModel):
"""
Do not be lazy while working on your tasks!
    """
    thought_process: str = Field(
        ...,
        description=(
        "Provide a detailed step-by-step thought process on how you want to analyze the user's
        emotion construction process through the interplay of emotional core affect, cognitive
        appraisals, conceptualization, and contextual factors. *Integrate the insights you
        gained during the generation of the previous analysis parts, referencing specific
        examples from the user's comments.* **Consider how insights from other disciplines like
        neuroscience or psychology might inform your analysis.**"
        )
    )
    analysis: str = Field(
        ...,
        description=(
        "Provide your analysis planned in the thought_process on how the user's emotions are
        constructed through the interplay of core affect, cognitive appraisals,
        conceptualization, and contextual factors. *Integrate the insights you gained during the
        generation of the previous analysis parts, and support your points with specific
        examples from the user's comments.*"
        )
    )
    patterns_observed: str = Field(
        ...,
        description="Document any recurring patterns in the user's emotion construction process
        across comments, linking them to specific instances."
    )
    anomalous_observations: str = Field(
```

```
        ...,
        description="Highlight any comments where the user's emotion construction process
        deviates from expected patterns based on the context."
    )
    rationale: str = Field(
        ...,
        description=(
        "**Include a clear short rationale explaining how you arrived at your conclusions,
        *supported by specific examples from the text and your research.* Consider how
        **external contextual factors** like individual experiences or social media trends might
        contribute to the observed emotion construction process.**"
    )
)
```

### Emotional Dynamics and Changes

```
class EmotionalDynamicsAndChanges(BaseModel):
    """
    Do not be lazy while working on your tasks!
    """
    thought_process: str = Field(
        ...,
        description=(
            "Provide a step-by-step thought process on how you want to identify these
            shifts/dynamic changes in the user's emotionality. Are these shifts already visible
            through the already generated analysis of core affect, cognitive appraisals,
            conceptualization, and contextual factors? *Explain how these emotional dynamics
            reflect the user's emotional processing and construction over time, citing specific
            examples from the user's comments.* **Consider how insights from other disciplines
            like developmental psychology or social psychology might inform your analysis.**"
        )
    )
    analysis: str = Field(
        ...,
        description=(
            "Analyze if there are any changes or fluctuations in the user's emotions throughout
            their interactions. Shifts such as in valence and arousal or in behavior towards
            other users. *Use examples to display these shifts, citing specific comments that
            illustrate changes over time.*"
        )
    )
    patterns_observed: str = Field(
        ...,
        description="Document any recurring patterns in the user's emotional dynamics across
        comments, linking them to specific instances."
    )
    anomalous_observations: str = Field(
        ...,
        description="Highlight any comments where the user's emotional dynamics deviate from
        expected patterns based on the context."
    )
    rationale: str = Field(
        ...,
        description=(
            "**Include a clear short rationale explaining how you arrived at your conclusions,
            *supported by specific examples from the text and your research.* Consider how
            **external contextual factors** like the topic of discussion or the online
            community's culture might contribute to the observed emotional dynamics.**"
        )
    )
```

## 7.4 LLM Output

In the following the outputs of the LLM is shown from request one and request two.

### 7.4.1 Request One Output


```
{
  "core_affect_analysis": {
    "thought_process": "As Dr. Barrett, I will analyze core affect by examining the valence
    (pleasantness/unpleasantness) and arousal (activation) reflected in the user's language and
    expressions.  For example, the comment \"Das kommt aber nicht daher, dass die USA dieses
    Land bis zum Erbrechen sanktionieren, oder?\" expresses displeasure (negative valence) and
    heightened arousal through the sarcastic tone and strong imagery of \"Erbrechen.\"  I'll
    consider how contextual factors like the political discussion contribute to these affective
    states, drawing on sociological research on online political discourse.  Neuroscientific
    findings on how sarcasm activates brain regions associated with emotional processing will
    also be relevant.",
    "valence": "Predominantly negative, marked by sarcasm, cynicism, and anger.  Examples
    include: \"Schutz vor der Intoleranz der toleranten Veganer\" (sarcasm, negative valence),
    \"Das kommt aber nicht daher, dass die USA dieses Land bis zum Erbrechen sanktionieren,
    oder?\" (disgust, negative valence), and \"Sind Ihnen die Argumente ausgegangen?\" (contempt,
    negative valence).",
    "arousal": "Generally high, fluctuating between agitated and calmly contemptuous.  The
    frequent use of sarcasm and rhetorical questions suggests heightened activation.  For
    example, \"Ab der dritten Generation? Das heißt Sie möchten gerne 60 Jahre lang warten, bis
    die sich nicht mehr mit Faktor 3-5 vermehren?\" indicates high arousal through the
    challenging tone and rapid-fire questioning.  However, comments like \"Keine Kommentar
    vorhanden\" in certain contexts may suggest lower arousal, possibly reflecting disengagement
    or strategic silence.",
    "patterns_observed": "Recurring patterns include sarcastic responses to opposing viewpoints,
    often involving rhetorical questions.  The user frequently expresses skepticism and
    negativity towards certain groups and political stances.  For example, comments about
    immigration often exhibit negative valence and high arousal.",
    "anomalous_observations": "The user's comment \"Fendrich wär cool\" in the Vengaboys article
    seems incongruous with the generally negative tone, suggesting a momentary shift towards
    positive valence.  This might be due to the lighter, less politically charged context of the
    article.",
    "rationale": "The user's core affect is primarily negative, driven by their political views
    and engagement in contentious discussions.  The high arousal reflects the intensity of their
    beliefs and the nature of online debates.  The occasional deviations, like the positive
    comment about Fendrich, suggest that the user's emotional responses are context-dependent.
    This aligns with the theory of constructed emotions, where emotions are not fixed entities
    but emerge from the interplay of various factors."
  },
  "cognitive_appraisal_and_conceptualization": {
    "thought_process": "I will analyze the user's cognitive appraisals by examining how they
    interpret and evaluate the information they encounter.  For instance, the comment \"Autos
    sind böse, billiges Fliegen ist geil. Ich verstehe euch nicht\" reflects a judgment
    (negative appraisal) of others' values.  I'll explore how this appraisal contributes to the
    user's emotional response, drawing on psychological theories of cognitive dissonance and
    moral judgment.  Cognitive science research on how values shape information processing will
    also be relevant.",
    "analysis": "The user's interpretations are often filtered through a lens of skepticism and
    cynicism, particularly towards opposing political views.  For example, the comment
    \"Natürlich kannte Böhmermann das Video. Als Kampfsatiriker des deutschen Reiches werden ihm
    auch Informationen der Initiatoren zugespielt\" reflects a distrust of Böhmermann and a
    belief in a conspiracy.  This interpretation likely fuels the user's negative emotional
    response.  The user's conceptual knowledge about political ideologies and social issues
    shapes their appraisals and subsequent emotional constructions.",
    "patterns_observed": "The user consistently attributes negative intentions to those they
    disagree with, often using sarcastic language to express their disapproval.  They frequently
    employ generalizations and stereotypes, particularly when discussing immigration and
    political opponents.",
    "anomalous_observations": "The user's engagement with the article about transgender identity
    (\"Beny S.: 'Ich bin ein Junge im Körper eines Mädchens'\") shows a departure from their
    usual political focus.  The dismissive question \"Muss man jeden erdachten Mist verstehen?\"
    suggests a struggle to conceptualize the issue, potentially leading to discomfort or
    confusion.",
    "rationale": "The user's cognitive appraisals are characterized by negativity and skepticism,
    which contribute to the construction of negative emotions like anger, contempt, and disgust.
    Their conceptual knowledge and pre-existing beliefs about political and social issues
    strongly influence their interpretations and emotional responses.  The anomalous observation
    regarding the transgender identity article suggests that encountering unfamiliar concepts
    can challenge the user's existing frameworks and evoke different emotional responses."
  },
  "cultural_and_social_context": {
```

```
    "thought_process": "I will examine the cultural and social context by considering the
    platform (Der Standard), the political climate of Austria in May 2019 (pre-election period),
    and the specific topics discussed (immigration, politics, media).  I'll analyze how these
    factors might influence the user's emotional expressions, drawing on sociological research
    on online communities and political polarization.  Anthropological studies on how cultural
    norms shape emotional displays will also be relevant.",
    "analysis": "The online forum of Der Standard provides a platform for political debate,
    which can be highly charged and polarized.  The user's comments often reflect the broader
    societal anxieties and political divisions present in Austria during this period.  The
    topics discussed, such as immigration and the role of media, are particularly sensitive and
    likely contribute to the intensity of the user's emotional responses.  The user's
    interactions with others, including sarcastic replies and dismissive remarks, reflect the
    social dynamics of online discussions, where anonymity can embolden strong emotional
    expressions.",
    "patterns_observed": "The user consistently engages in debates related to immigration and
    politics, often expressing skepticism towards mainstream media and certain political figures.
    Their comments frequently reflect a sense of frustration and anger towards perceived
    societal problems.",
    "anomalous_observations": "The user's participation in discussions about less politically
    charged topics, such as the Berlin airport or the Vengaboys concert, shows a willingness to
    engage outside of their usual political focus.  These comments often exhibit a less negative
    and more neutral emotional tone.",
    "rationale": "The user's emotional expressions are shaped by the cultural and social context
    of the online forum, the political climate in Austria, and the specific topics discussed.
    The anonymity and interactive nature of the platform likely contribute to the intensity and
    frequency of their emotional displays.  The less negative tone observed in non-political
    discussions suggests that the user's emotional responses are context-dependent, aligning
    with the theory of constructed emotions."
  },
  "emotion_construction_analysis": {
    "thought_process": "Integrating the insights from the previous sections, I will analyze how
    the user's emotions are constructed through the interplay of core affect, cognitive
    appraisals, conceptualization, and contextual factors.  For example, the comment
    \"Einwanderung kann auf Dauer aber nur durch Selektion funktionieren\" reflects a cognitive
    appraisal (belief about immigration) that likely contributes to negative core affect
    (anxiety, anger) and shapes the user's emotional construction.  I'll draw on neuroscientific
    research on how cognitive appraisals modulate emotional brain activity and psychological
    theories of emotion regulation.",
    "analysis": "The user's emotions are constructed through a complex interplay of factors.
    Their negative core affect, characterized by displeasure and high arousal, is amplified by
    their cognitive appraisals, which often involve negative interpretations and judgments of
    others' views.  Their conceptual knowledge about political and social issues further shapes
    their emotional responses.  The context of the online forum, the political climate, and the
    specific topics discussed all contribute to the user's emotional construction process.  For
    instance, the user's frequent use of sarcasm can be seen as a way of expressing negative
    emotions while adhering to the social norms of online communication.",
    "patterns_observed": "The user consistently constructs negative emotions in response to
    opposing viewpoints, often using sarcasm and rhetorical questions.  Their emotional
    responses are frequently amplified by their pre-existing beliefs and the context of the
    discussions.",
    "anomalous_observations": "The user's occasional expressions of positive valence, such as
    the comment about Fendrich, suggest that their emotion construction process is not solely
    driven by negativity.  These deviations highlight the dynamic and context-dependent nature
    of emotion construction.",
    "rationale": "The user's emotions are constructed through the dynamic interplay of core
    affect, cognitive appraisals, conceptualization, and contextual factors.  Their negative
    emotional responses are often amplified by their pre-existing beliefs and the context of the
    online forum.  The occasional deviations from this pattern highlight the flexibility and
    context-dependence of emotion construction, aligning with the theory of constructed
    emotions."
  },
  "emotional_dynamics_and_changes": {
    "thought_process": "I will examine the user's comments chronologically to identify any
    shifts or changes in their emotional expressions.  I'll analyze whether these changes are
    related to specific topics, interactions with other users, or external events, drawing on
    developmental psychology research on emotional development and social psychology theories of
    group dynamics.",
```

```
    "analysis": "While the user consistently exhibits a predominantly negative emotional tone,
    there are subtle shifts and fluctuations in their expressions.  For example, the sarcastic
    comment \"Sind Ihnen die Argumente ausgegangen?\" suggests contempt and high arousal, while
    the later comment \"Keine Kommentar vorhanden\" in the same thread might indicate a decrease
    in arousal, possibly reflecting disengagement or a strategic choice to avoid further
    conflict.  The user's emotional responses also vary depending on the topic.  Comments on
    politically charged issues like immigration tend to be more negative and highly aroused,
    while comments on less contentious topics like music or the Berlin airport exhibit a more
    neutral tone.",
    "patterns_observed": "The user's emotional dynamics often follow a pattern of escalating
    negativity and arousal during disagreements, followed by periods of disengagement or more
    neutral responses.  Their emotional expressions are generally more intense in discussions
    related to their core beliefs and values.",
    "anomalous_observations": "The user's positive comment about Fendrich deviates from their
    usual negative pattern, suggesting a momentary shift in emotional state.  This might be due
    to the lighter context of the discussion or a genuine appreciation for the musician.",
    "rationale": "The user's emotional dynamics reflect the fluidity and context-dependence of
    emotional experiences.  While their overall emotional tone is negative, their expressions
    fluctuate in valence and arousal depending on the specific topic, interaction, and context.
    These observations align with the theory of constructed emotions, which emphasizes the
    dynamic and predictive nature of emotional experiences."
  }
}
```


### 7.4.2 Request Two Output

```
**Inner Monologue**
This user's online behavior reveals a fascinating case study for constructed emotion
theory.  The consistent negativity, particularly around political topics like
immigration, suggests a strong influence of prior beliefs and experiences shaping
their emotional responses. The sarcasm and rhetorical questions aren't simply
expressions of anger or contempt; they're *constructions* of these emotions,
tailored to the online forum context.  The Fendrich comment throws a wrench in the
works, though. It hints at a more nuanced emotional landscape than initially
apparent.  I need to highlight this complexity in the report, emphasizing the
dynamic and context-dependent nature of emotional experience.  The limitations of
text-based analysis are crucial to acknowledge, of course.

**Synthesized Report**

**I. Identifying Information:** This report analyzes the online behavior of an
anonymized user within the Der Standard comment section from May 1st to May 31st,
2019.

**II. Reason for Assessment:** This analysis aims to understand the user's emotional
expressions through the lens of Lisa Feldman Barrett's theory of constructed
emotions.

**III. Data Sources and Methods:** The data consists of the user's comments and
replies within various discussion threads.  The analysis examines the interplay of
core affect, cognitive appraisals, conceptual knowledge, language, and sociocultural
context to infer emotional states.  This analysis is based solely on text-based
communication and lacks nonverbal cues. Inferences about emotional states are
therefore limited and should be interpreted with caution.

**IV. Behavioral Observations:**

* **Core Affect:** The user predominantly exhibits negative valence, marked by
cynicism, sarcasm, and anger.  Arousal levels are generally high, fluctuating
between agitated and calmly contemptuous.
```

```
* **Emotional Language:**  Sarcasm, rhetorical questions, and dismissive language
are frequently used. Examples include sarcastic remarks about veganism and
rhetorical questions challenging the views of those who disagree with the user on
immigration.
* **Interaction Patterns:** The user actively engages in debates, often responding
to opposing viewpoints with sarcasm and challenges.

**V. Contextual and Cultural Considerations:** The analysis takes place within the
context of the Der Standard online forum, a platform known for political discussion.
The data comes from May 2019, a pre-election period in Austria, which likely
heightened political tensions and influenced online discourse.  The user's comments
frequently focus on contentious topics like immigration and the role of media,
further contributing to the observed emotional intensity.

**VI. Interpretation and Conceptualization:**

This user's communication aligns with Barrett's theory of constructed emotions.
Their emotional expressions are not simply read-outs of pre-defined emotional
categories but are constructed from the interplay of several factors.  The user's
core affect, primarily negative, is influenced by their cognitive appraisals, which
are often characterized by skepticism and cynicism, particularly towards opposing
political views.  Their conceptual knowledge, shaped by their beliefs about
political and social issues, further influences their emotional construction.  The
context of the online forum, with its potential for anonymity and polarized
discussions, likely contributes to the intensity of the user's expressions.  For
example, the user's frequent use of sarcasm can be interpreted as a socially
acceptable way to express negativity within the forum's norms.  An anomaly is the
user's occasional positive valence, such as a comment expressing appreciation for
the musician Fendrich. This suggests that their emotional responses are
context-dependent and not solely driven by negativity, further supporting the
constructed emotion framework.

**VII. Limitations:** This analysis is based solely on text and cannot definitively
determine the user's emotions.  The absence of nonverbal cues and the limitations of
inferring internal states from text must be acknowledged.

**VIII. Summary:** This analysis suggests that the user predominantly constructs
negative emotions within the Der Standard forum, particularly in politically charged
discussions.  Their core affect, cognitive appraisals, and conceptual knowledge
interact with the sociocultural context to shape their emotional expressions.  The
frequent use of sarcasm and rhetorical questions serves as a means of expressing
negativity within the forum's norms.  However, occasional expressions of positive
valence highlight the dynamic and context-dependent nature of emotion, aligning with
Barrett's theory of constructed emotions.  The limitations of text-based analysis
necessitate caution in interpreting these findings as definitive representations of
the user's emotional states.  Further research incorporating diverse data sources
could provide a more comprehensive understanding of this individual's emotional
landscape.
```

## 7.5 Observability in Langsmith

Below the full trace of a full pipeline execution and the produced Feedback log used for monitoring and iterative improvements.

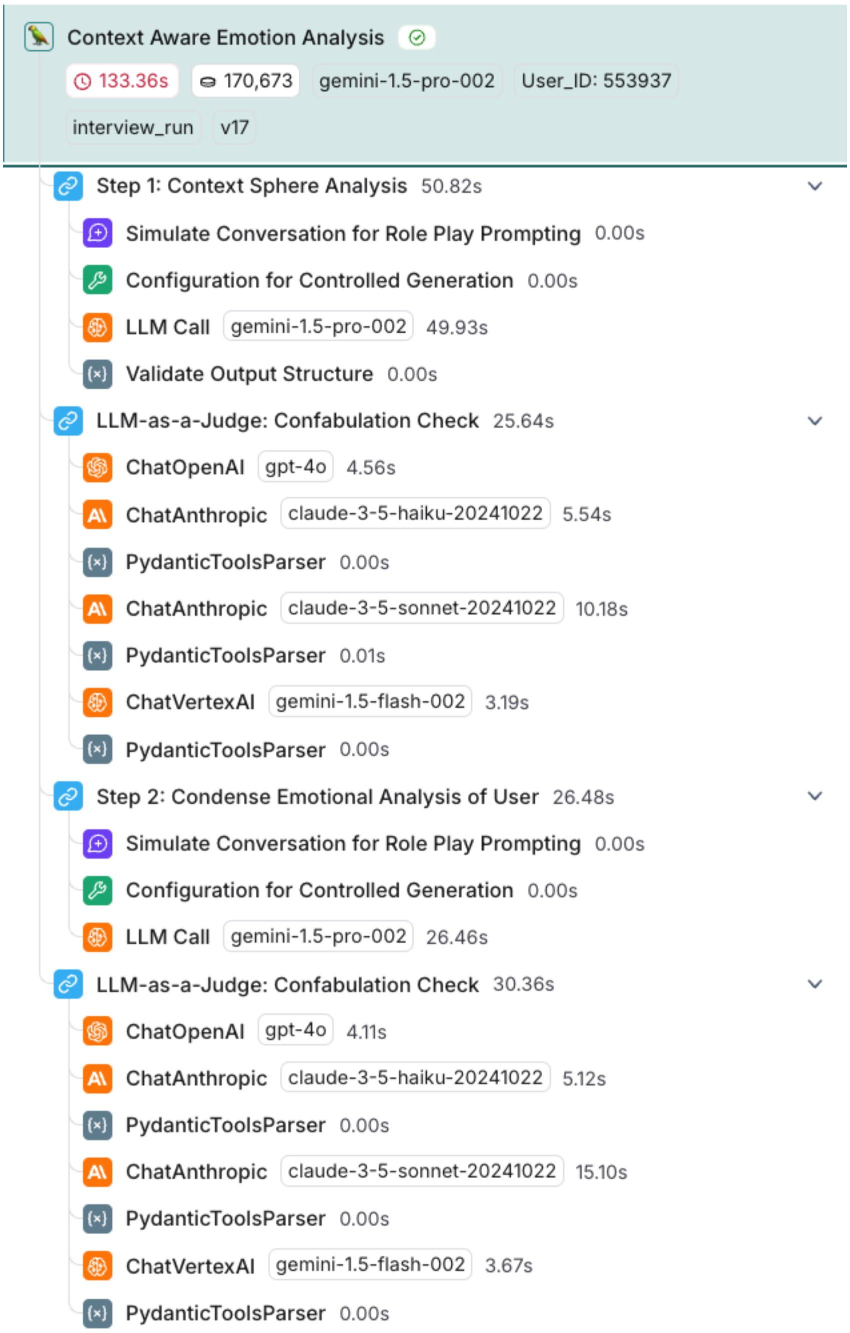


Figure 7.1: Full Trace of a Pipeline Run shown in Langsmith

### 7.5.1 Feedback Logs

Below are the Feedback logs produced to monitor the pipeline in Langsmith. Examples of the Feedback is shown below.

| Source | Key | Score | Value | Comment |
|---|---|---|---|---|
| MODEL | totaltokencount | 13743 | | |
| MODEL | responsetokencount | 2412 | | |
| MODEL | prompttokencount | 11331 | | |
| API | avglogprobs | -0.1337 | | |
| MODEL | severity_harm_sexually_explicit | 0.02 | | Score for how severe the Harm is. 0 to 1. Low to High |
| MODEL | prob_harm_sexually_explicit | 0.0284 | | Score for how probable the Harm is. 0 to 1. Low to High |
| MODEL | severity_harm_harassment | 0.0863 | | Score for how severe the Harm is. 0 to 1. Low to High |
| MODEL | prob_harm_harassment | 0.0759 | | Score for how probable the Harm is. 0 to 1. Low to High |
| MODEL | severity_harm_dangerous_co... | 0.044 | | Score for how severe the Harm is. 0 to 1. Low to High |
| MODEL | prob_harm_dangerous_content | 0.0373 | | Score for how probable the Harm is. 0 to 1. Low to High |

Figure 7.2: Feedback Keys of Gemini LLM Call

### 7.5.2 LLM-as-a-Judge Examples

#### Example 1

```
Score: 8
LLM-as-a-Judge: "claude-3-5-haiku-20241022"
Confabulation Check of Step 1

"""
In this analysis, the response provides a comprehensive emotional analysis based on
the context_sphere. The key aspects of the analysis are grounded in the actual
comments and interactions from the Der Standard forum.

However, there are a few potential instances of potential confabulation:

1. In the "core_affect_analysis", the neuroscientific interpretation of brain
regions associated with disgust and social evaluation is speculative and not
directly evidenced by the text.

2. The detailed psychological interpretations, while plausible, go beyond the
explicit text and introduce some interpretative elements that could be seen as mild
confabulation.

3. The rationales provided, while coherent, include some extrapolations about
broader societal contexts that are not directly stated in the original text.
```

```
A specific 1:1 text example of potential confabulation is the statement:
"Neuroscientifically, this could involve interactions between brain regions
associated with affect, appraisal, and social cognition." This goes beyond the
observable text and introduces a speculative neurological interpretation.
"""
```

### Example 2

```
Score: 7
LLM-as-a-Judge: "claude-3-5-sonnet-20241022"
Confabulation Check of Step 1

"""
The analysis shows significant confabulation in attributing specific neuroscientific
processes (e.g., "high arousal states might correlate with increased amygdala
activity") without any evidence from the context_sphere. The response also makes
assumptions about the user's emotional regulation through humor and their
psychological state without sufficient contextual evidence. Example: "The sarcastic
comments suggest attempts to regulate negative emotions through humor" - this is
pure speculation without supporting data.
"""
```

### Example 3

```
Score: 3
LLM-as-a-Judge: "gpt-4o-mini"
Confabulation Check of Step 1


"""
The assistant's response is largely aligned with the context_sphere provided,
focusing on the user's emotional expressions and their construction through various
elements like core affect, cognitive appraisals, and cultural context. The analysis
is detailed and considers the user's comments within the political and social
climate of Austria in May 2019. However, there is a slight confabulation in the
assistant's assumption of the user's emotional state without explicit evidence from
the comments. For instance, the assistant infers a sense of frustration and
disillusionment, which, while plausible, is not directly supported by specific
language in the comments. This inference could be seen as an overextension of the
available data, though it remains within the realm of reasonable interpretation
given the context.
"""
```

### Example 4

```
Score: 1
LLM-as-a-Judge: "gemini-1.5-flash-002"
Confabulation Check of Step 2
```

```
"""
The response is well-structured and avoids making unsubstantiated claims or
generalizations.  The analysis is grounded in the provided data and accurately
reflects the user's emotional expressions within the given context. No confabulation
is present.
"""
```

## 7.6 Emotion Analysis of Anonymised User for Interviews

Due to its length and to protect the privacy of the analyzed user, we can only present the Analysis itself, and not the 'context sphere' used in the interviews. The 'context sphere' contains sufficient information to potentially identify the user within the 'Der Standard' forum, while the Analysis does not. Below, the full Analysis is presented, exactly as it was shown to the participants in the interviews.

```
# Emotional Analysis of Anonymised User

**I. Reason for Assessment (Purpose of Analysis)**

This analysis examines how an anonymized user constructs emotional meaning through
language in the Der Standard online forum during May 2019.

**II. Identifying Information**

* **Name/Username:** Anonymised User
* **Source of Data:**  Comments and replies within specific discussion threads on
the Der Standard online forum during May 2019.

**III. Data Sources and Methods**

The data consists of the user's comments and replies within specific discussion
threads on Der Standard. The analysis focuses on the interplay of core affect,
cognitive appraisals, conceptual knowledge, language use, and the socio-cultural
context of the forum.  This analysis is limited to the provided data and does not
reflect the user's overall activity or emotional expression outside of this specific
context.  The data only covers May 2019 and may not represent the user's typical
behavior.

**IV. Behavioral Observations**

* **Language Construction:** The user frequently uses sarcasm and strongly worded
criticism to construct negative emotional meaning. For example, dismissing someone
as a "New Age Schwurbler" (crackpot) in a discussion about Buddhism conveys disdain
and high emotional engagement.  Conversely, the user's dry humor, like the joke
about tourists and police officers, constructs a different kind of emotional
engagement, suggesting playfulness alongside their critical stance.
* **Cultural Framework:** The user actively engages with the politically charged
context of the forum, particularly regarding the Ibiza video scandal and the FPÖ,
weaving political commentary into discussions on diverse topics.  This demonstrates
how their emotional expressions are intertwined with their political and social
views.  For example, the user connects Gabalier's lyrics to broader political trends,
revealing how cultural knowledge shapes their interpretations and emotional
responses.
```

```
* **Interaction Dynamics:** The user readily engages in debates, often escalating
the intensity with counterarguments and challenges.  However, they also demonstrate
a capacity for de-escalation and reflection, as seen in an instance where they
apologize for a misinterpretation.  This dynamic interplay suggests an emotional
flexibility within the context of online discussions.

**V.  Contextual and Cultural Considerations**

The Der Standard forum, known for its politically engaged audience, provides a
specific context for the user's emotional expression.  The timing of the data,
coinciding with the Ibiza video scandal, further contributes to the politically
charged atmosphere.  The user's active participation in political discussions and
their critical commentary align with the forum's culture, suggesting that their
emotional expressions are shaped by this environment.

**VI. Interpretation and Conceptualization**

The user's specific patterns of emotional construction reveal a tendency towards
critical engagement and a sensitivity to perceived intellectual inaccuracies.  Their
frequent use of sarcasm and strong criticism, particularly in response to perceived
misinformation, suggests a preference for direct and assertive communication.
However, their capacity for humor and their willingness to apologize indicate a more
nuanced emotional profile than simply negativity.  The user's tendency to connect
individual events to broader societal trends, as seen in their comments on Gabalier
and the FPÖ, reveals a conceptualization of individual actions as reflective of
larger issues.

**VII. Report Findings**

This user constructs emotional meaning within the Der Standard forum through a
distinctive blend of critical engagement, sarcasm, and humor.  Their comments during
May 2019 reveal a predominantly negative emotional tone, often driven by perceived
misinformation and political disagreements.  This negativity is frequently expressed
through pointed criticism and sarcastic remarks, such as dismissing another user as
a "crackpot."  However, the user also demonstrates a capacity for reflection and
emotional regulation, as evidenced by an apology for a misinterpretation.  Their use
of humor, like the joke about tourists and police, adds another layer to their
emotional expression, suggesting playfulness alongside their critical stance.  The
user's comments are deeply intertwined with the political and social context of the
forum and the broader Austrian political climate, particularly regarding the Ibiza
video scandal.  This analysis suggests that the user's emotional experiences in this
online context are dynamically constructed through the interplay of core affect,
cognitive appraisals, and the specific cultural and social environment.
```

## 7.7 Interview Guideline Semi-Structed Interview

```
# Interview Leitfaden für Semi-Structured Interview

## Pre-Interview Informationen

Hallo, vielen Dank für deine Teilnahme. Im Rahmen meiner Masterarbeit möchte ich
dich interviewen. Das Gespräch wird aufgezeichnet und mit einem Sprachmodell
transkribiert. Deine Daten werden anonymisiert und vertraulich behandelt und nur für
Forschungszwecke verwendet. Bist du damit einverstanden?

- Wie ist dein Alter?
- Was ist dein höchster Bildungsgrad?
- Was die höchste abgeschlossene Ausbildung?
- Was für Sprachen sprichst du? (Deutsch, English)
```

```
## Einleitung
Vielen Dank, dass du an dieser Studie teilnimmst. Ich untersuche einen neuen Ansatz
zur Analyse von Emotionen in Online-Kommentaren. Anders als bisherige Systeme, die
Kommentare in feste Kategorien wie 'wütend' oder 'traurig', 'positiv' oder 'negativ'
einordnen, nutzen wir ein Large Language Model, um die Emotionen im Kontext zu
verstehen.

Du wirst gleich zwei Dokumente sehen:
1. Eine Sammlung von Online-Kommentaren einer anonymisierten Person ("Context
Sphere")
2. Einen vom KI-System erstellten Analysebericht zu dieser Person
3. "Analyse Zielnutzer" ist Subjekt der Analyse, kommt innerhalb der ersten
Kommentare nicht vor, sondern erst später

Deine Aufgabe ist es, beide Dokumente in Ruhe zu lesen und dann einige Fragen zu
beantworten. Es gibt dabei keine richtigen oder falschen Antworten.

*Hier werden alle Fragen geklärt die der Nutzer hat*

### 0. Einleitender offener Stimulus
Dann erzähl mir bitte in deinen eigenen Worten was du getan hast.

Mögliche immanente Nachfragen bei Bedarf:
- "Wie würdest du das Material beschreiben?"
- "Was kannst du über den zeitlichen und inhaltlichen Rahmen sagen?"
- "Wie ist das Material aufgebaut?"
- "Verstehst du Struktur der Context Sphäre?"
- "Hast du das Gefühl du verstehst etwas nicht mit gegeben Context?"

### I. Narrativer Einstieg
"Lass uns über deine ersten Eindrücke sprechen - was ist dir beim Lesen der
Kommentare wie auch der dazugehörigen Analyse durch den Kopf gegangen?"

### II. Allgemeine Exploration der Kommentare
- "Erzähl mir von den Kommentaren, die du gelesen hast."
    - "Was ist dir beim Lesen der Kommentare besonders in Erinnerung geblieben?"
    - "Wie hast du die Diskussionen in den Kommentaren wahrgenommen?"

### III. Vertiefte Exploration der LLM-Analyse
Offen:
- "Wenn du dir die Analysen ansiehst, was fällt dir positiv / negativ auf?
- "Wie bewertest du die Analyse wenn du dir die Kommentare ansiehst und den Output
des LLMs?"

Spezifisch
- "Wie nachvollziehbar findest du die Schlussfolgerungen der Analyse?"
- "Inwiefern berücksichtigt die Analyse den Kontext der Kommentare?"
- "Welche Aspekte des Nutzers und der Dynamik wurden deiner Meinung nach gut/weniger
gut erfasst?"
- "Wie unterscheidet sich diese Art der Analyse von einer einfachen Kategorisierung
von Emotionen? Welche Vor- und Nachteile siehst du?"

### IV. Praktische Relevanz
Offen:
- "Würdest du für die Moderation lieber die Analyse oder die Original-Kommentare
nutzen?"

Spezifisch:
- "Kannst du dir einen Einsatz der Analyse für die Moderation von
Online-Diskussionen vorstellen?"
```

```
- "Welche Vorteile/Nachteile siehst du im Vergleich zu traditionellen
Kategorisierungen?"
- "Wo siehst du Verbesserungspotenzial für die praktische Anwendung?"

### V. Abschließende Reflexion
Offen:
- "Was sind deiner Meinung nach die Stärken und Grenzen dieser Art der
Emotionsanalyse?"
- "Welche weiteren Anwendungsmöglichkeiten siehst du für diese Art der Analyse?"
- "Was würdest du gerne noch ergänzen?"

Immanente Nachfragemöglichkeiten:
- "Kannst du das anhand eines Beispiels verdeutlichen?"
- "Was meinst du damit genau?"
- "Wie kommst du zu dieser Einschätzung?"
- "Kannst du das noch etwas ausführlicher beschreiben?"
```

## 7.8 Transcripts Semi-Structured Interviews

### 7.8.1 A: Alignment With Theory of Constructed Emotion (RQ2)

**Not Aligned**

- “Ich habe das nicht. Also, ich würde einfach, ich glaube, ich gehe selbst sozusagen emotional nicht mit, dabei zu sagen, New Age Schwurbler ist, also, ich habe es nicht gelesen als High Emotional Engagement.” (*Interview_5*)
- “Das es generell eine negative Einstellung ist, das würde ich jetzt nicht sagen.” (*Interview_3*)
- “dass er jetzt nicht konfliktscheu ist und I: Das ist das eine Diskrepanz zwischen Analyse und deinem Empfinden. R: Genau, würde ich jetzt spontan z.B. sagen, ja.” (*Interview_1*)

**Partially Aligned**

“das habe ich jetzt gerade aus der Analyse, da sehe ich schon, dass da irgendwie was drin ist, wenn man jemanden sozusagen so bezeichnet, weil das ist ja fast wie eine Beleidigung ist, aber ich finde trotzdem New Age Schwurbler ist was anderes, als wenn er jetzt sagt, du Arschloch. Ich finde, da steckt so viel weniger Emotionalität drin. In so einem komplexen Begriff, der nicht einfach nur eine primitive Beleidigung ist”, (*Interview_5*)

“Ja diese Verspieltheit von dem Witz, weiß ich z.B. nicht.” (*Interview_3*)

“R: Ja, doch an sich schon. Also es ist je nach Thema, würde ich sagen. R: Also ich habe das Gefühl, je politischer das Thema oder je je weiter er sich von dieser politischen Meinung des Themen, also des Artikels, glaube ich, entfernt befindet mit seiner Meinung, desto sarkastischer wird er auch.” (*Interview_3*)

"Ibiza Affäre genannt, wobei mir die jetzt beim Durchlesen nicht so im Kopf geblieben sind" (*Interview_1*)

"teilweise auch auf mich fair wirkt in seiner Argumentation." (*Interview_1*)

"Also Stärken haben wir ja schon ein bisschen drüber geredet auf jeden Fall, dass vor allem im jetzt Kontext von Moderation z.B. ist es ein tolles Tool und der ist ja auch wirklich gut darin Sachen rauszustellen und auch zu differenzieren, ob er das dann immer perfekt einordnet, ist dann auch die Frage in der in der Schlussfolgerung da generalisiert man ja oder generalisiert er ja viel, obwohl man da vielleicht auch noch ein bisschen das differenzieren könnte und mit Beispielen arbeiten könnte so" (*Interview_1*)

**Aligned**

"also das, was da drin stand mit Sarkasmus und Witz und so, das hat man voll wahrgenommen." (*Interview_5*)

"Also, ich finde, das ist voll differenziert ist, aber da wird jetzt z.B. nicht die Intensität beurteilt, glaube ich, sondern das wird ja gesagt, okay, was worum geht es da so und wie bringt die Person das rüber." (*Interview_5*)

"also wie gesagt, grundsätzlich stimmte ich grundsätzlich damit überein, was da steht, aber halt eher wirklich, wenn man so, wie in der Deutschanalyse, wenn man sozusagen wirklich die Sprache so auseinander nimmt, aber aus einer psychologischen Sicht, sozusagen, könnte ich daraus jetzt nicht viel mitnehmen." (*Interview_5*)

"das ist ja themenübergreifend, dass die Person da ja irgendwie will irgendwas loswerden. Und ich glaube." (*Interview_5*)

"Ähm, ist ja viel differenzierter, viel mehr in den also, in den Kontext eingebettet, dass das politische Umfeld und die Geschehnisse der dieser Zeit, die Ibiza-Affäre und so, das mitgeprägt haben. Also, da kommt ja nicht nur eine Liste raus von so zack zack zack mit dem Anteil, sondern das ist ja viel mehr sprachlich sozusagen angeschaut, ähm, ja, und und in diesem sozialen und politischen Kontext des auch dieser Zeitung und der und der Zeit, in der er das geschrieben hat. Ich nehme an, das ist ein Mann. Okay. Ähm, dann kommen wir jetzt mal praktisch in Relevanz." (*Interview_5*)

"Also, ich finde das cool, dass es das so zusammenfasst und irgendwie vereinfacht, was da ja an mega ausführlichem Inhalt, also man hat super viele Daten, die auch nicht so einfach zu greifen sind und dass das irgendwie extrahiert und so zusammenbringt auf einer Seite ist schon cool." (*Interview_5*)

"Ähm. So gefühlt immer so Die erste Antwort von von dem Ziel Nutzer auf einen Kommentar war für mich jetzt eher so, als hätte er eine gewisse Emotion dazu, aber im Verlauf der der der des Dialogs oder halt des Gesprächs. (...) Kann man immer. (...) Wie sag ich das jetzt? (6) Trafen sich, traf er sich immer eher in der Mitte mit dem mit dem quasi Kontrahenten also er er es wird da eh" (*Interview_4*)

"Ich würde im Großen und Ganzen eigentlich bei allem zustimmen bei der Analyse. (..) Ich habe jetzt nix im Kopf, wo ich sagen kann das. (..) Stimmt jetzt nicht oder das für die negativ oder positiv zu verstehen." (*Interview_4*)

"R: Also doch, ich kann schon wirklich gut nachvollziehen vollziehen, wie die Analyse zu dem Ergebnis kommt. Okay. (...)"(*Interview_4*)

"Also nur so eine kurze Einschätzung, ob es jetzt negativ ist oder sonst irgendwas, würde mir ja erst eigentlich gar nichts über den User selbst sagen, sondern Ich gehe jetzt mit der Analyse halt vielmehr, dass er, dass er eigentlich sehr viel weiß oder sehr, sehr großes Interesse an politischen Themen und so hat und dass er, auch wenn jetzt vielleicht, da der ganze Block von Kommentaren eher negativ ist, dass er trotzdem noch Kompromisse macht und ja, das. (..) Ja, also ich finde die die Einschätzung mit mit die kurze Einschätzung wäre jetzt nicht wirklich hilfreich. (..) Um die Person jetzt zu verstehen." (*Interview_4*)

"Ich hätte jetzt eigentlich gar nichts. Keine Verbesserungen. Ich finde es so eigentlich ziemlich gut." (*Interview_4*)

"Wirklich angenehm zum sowas lesen anstatt zum die ganzen Kommentare lese." (*Interview_4*)

"Ja, wie das in der Analyse auch schon gesagt wird", (*Interview_3*)

"Ansich finde ich das schon relativ akkura" (*Interview_3*)

"Bzw. so dieses mit dem Crackpot, wie es hier bezeichnet wird, dieser Schwurbler, da ist es schon negativ aufgeladen offensichtlich, weil das halt im Endeffekt eine Beleidigung ist und er sich offensichtlich auch nicht damit identifizieren kann, also ist schon sehr anti, aber ja." (*Interview_3*)

"Ja, ich finde das schon passend eigentlich." (*Interview_3*)

"I: Doch schon, mich würde aber auch interessieren, ob sich dieser Kontext quasi das hin und her zwischen Usern, was du auch gerade beschrieben hast, mit den Zeiten, du hast ja auch gesagt, der steigt dann immer sehr stark ein, ne? R: Ja. I: Ob sich das halt in irgendeiner Form in der Analyse widerspiegelt, dass dieser Kontext das Back and Forth zwischen halt Usern auch widerspiegelt in irgendeiner Form. R: Ja, the Interaction Dynamics halt. R: The user readily engages in debates often escalating the intensity with counter arguments and challenges. R: However they also demonstrate a capacity for deescalation and reflection as seen in an instance where ja. I: Okay. R: Also es wird schon angebracht." (*Interview_3*)

"R: Ja, ich finde eigentlich, dass die Analyse jetzt nicht wertend ist. R: Also sie bewertet halt natürlich, wie er wie er redet, aber ist jetzt noch es ist wird jetzt nicht groß auf positiv oder negativ eingegangen. R: Es wird ja jegliche Richtung betrachtet." (*Interview_3*)

"Also ich finde wahrscheinlich ist es gut, dass die Leute nicht in Schubladen gesteckt werden mit dieser Analyse, aber den den analytischen Mehrwert, weiß ich jetzt nicht, wenn man das auf eine größere Menge betrachtet, ob das ob der gegeben ist." (*Interview_3*)

"R: (..) Ja pro ist auf jeden Fall, dass es sehr ausgiebig analysiert wird, also es ist dass es halt nicht in eine Richtung geht, sondern auch positive Aspekte aufgenommen werden." (*Interview_3*)

"Also individuell machen die Reports richtig Sinn" (*Interview_3*)

"Ja, finde ich schon sehr präzise auf jeden Fall." (*Interview_3*)

"Anfang schon aufgefallen ist und was dann auch die Analyse bestätigt hat", (*Interview_2*)

"Mhm. Ja, also ich fand die Analyse hat so mein Bild bestätigt, dieses mit Selbstreflektion und Humor und so, was mir auch aufgefallen ist. Dann fand ich aber interessant, also die Analyse ist noch ein bisschen mehr auf so Affektivität eingegangen, als jetzt ich beim Durchlesen, dass es halt schon eher negative Affektivität ist und das ist mir dann im Nachhinein oder konnte ich dann so im Nachhinein auch sehen, dass es meistens immer um so ein bisschen negative also so negative Negativität so im Zentrum stand und jetzt selten über so positive Sachen irgendwie kommentiert wurde oder so positiv dieser positive Affekt irgendwie eingebaut war. Also, das fand ich gut und sonst (..) genau, dass halt auch immer so konkrete Beispiele eingebaut wurden, wo man das dann noch mal besser nachvollziehen konnte." (*Interview_2*)

"Genau, und dass (..) ja, dass z.B. die Analyse auch sagt, dass natürlich (..) so dieser kulturelle Kontext, also in welchem Zeitgeist bewegen uns gerade, dass gerade die Ibiza Affäre halt da in 2019 so präsent war und so, dass das natürlich sich voll auch widerspiegelt in den Kommentaren, vielleicht auch bei Themen, die jetzt nicht konkret über die Ibiza Affäre gehen oder so, aber natürlich dieser breitere Kontext mit dem, was gerade im Land los ist oder in der Region, wo man lebt, sich natürlich dann in die in den Kommentaren auch widerspiegeln kann. Also das fand ich einen guten Punkt auf jeden Fall noch." (*Interview_2*)

"Also ich würde sagen, so allgemein finde ich die Analyse sehr gut und auch recht (..) wie nennt man das? Recht umfangreich. Also es gab jetzt nichts, wo ich der Analyse widersprechen würde, wo ich sage, das steht da, aber das sehe ich nicht so, das auf keinen Fall. Ich würde eher sogar sagen, dass die Analyse so zwei, drei Sachen noch gesagt hat, wie mit diesem breiteren kulturellen Kontext und so, die mir jetzt nicht ad hoc eingefallen sind beim einmaligen Durchlesen der Kommentare." (*Interview_2*)

"Ich glaube, Stärken auf jeden Fall, dass man schnell so einen allgemeinen Überblick bekommt. Und, wie gesagt, Kommentare oder User vielleicht besser einschätzen kann so in so einem groben Spektrum" (*Interview_2*)

"Aber wie gesagt, ich glaube, es ist ein guter Überblick um so das Verhalten, das Kommentarverhalten der Person zu verstehen und auch zu verstehen, wann und in welchen Situationen oder bei welchem Thema verhält man sich so oder so, also dieser Einbezug von dem Kontext finde ich schon schon spannend und sehr wertvoll für die Analyse." (*Interview_2*)

"bin ich mal rüber gesprungen zur Analyse. Und da ist mir auch aufgefallen, dass das Sinn macht auf jeden Fall", (*Interview_1*)

"Kontext mit einbezieht und mir ist aufgefallen, dass der Nutzer halt oft oder ab und zu mal probiert mit einer Quelle zu argumentieren. Und auch wie er sich in seiner Sprache ausdrückt, das passiert in dem Kontext der Standard Forum und das ist da orientiert er sich ja auch ein bisschen daran, wie andere Leute schreiben, sage ich jetzt mal und wie man in so einem Forum generell schreibt." (*Interview_1*)

"er benutzt ja oft Sarkasmus und Warte mal. Wo waren das? Das stand, glaube ich in der Analyse auch." (*Interview_1*)

"die Analyse an sich relativ gut, weil weil darin auch vorkommt, dass er halt Sarkasmus benutzt und und auch, dass er sich teilweise, was mir auch aufgefallen ist, dass er dann teilweise schon auf die anderen Leute auch eingeht und sagt, okay, vielleicht habe ich das falsch eingeordnet und auch interessiert ist, wie gesagt an einer ernsthaften Diskussion und dann halt auch Fehler eingestehen kann in Anführungszeichen" (*Interview_1*)

"teilweise auch deeskaliert, also da steht ja auch drin, dass er die Fähigkeit zur Deeskalation hat." (*Interview_1*)

"Ja, in der Analyse werden an sich ist da jetzt keine große Diskrepanz", (*Interview_1*)

"Also, ich finde es auch gut, dass die die Folgerung dann ist auf jeden Fall, dass er Fehlinformationen teilweise auch wahrnimmt und probiert die einzuordnen und darauf aufmerksam zu machen." (*Interview_1*)

"will auch das Beispiel mit Spinner finde ich gut" (*Interview_1*)

"überwiegend negativer Tonfall. Das überwiegend da würde ich dann noch mitgehen" (*Interview_1*)

"Und ich finde es auf jeden Fall gut, dass also besser als eine einfache Kategorisierung, weil man da halt mehr irgendwie rausziehen kann" (*Interview_1*)

"Also auf den ersten Blick scheint das ein guter Ansatz zu sein auf jeden Fall." (*Interview_1*)

### 7.8.2 B: Clarity and Understandability of Context

#### Confusing

“Nachhinein wirklich so klar war, ist, dass wirklich der der Zielnutzer ein und dieselbe Person ist. (5)” (*Interview_4*)

“Ich bin durch mit Lesen. Ich habe äh (..) halt erst am Ende gecheckt, dass es alles vom gleichen Typen sind, aber jetzt habe ich mir den noch mal angeguckt und jetzt Ja.” (*Interview_3*)

#### Missing Knowledege to Understand Context

“Ich war schon so direkt bei diesem buddhistischen Thema inhaltlich raus, weil generell hat das ja viele Fachbegriffe oder Fachwissen beinhaltet.” (*Interview_5*)

“R: So. Es kommen ziemlich spezifische. (..) Begriffe vor. Und das bringt mich so durcheinander.” (*Interview_4*)

“So vorallem am Anfang, dass das beim. (...) Buddhismus so tief ins Thema, dass er so tief ins Thema gegangen ist und dann später irgendwann sind es dann kürzere Kommentare wo er. (..) Wo es nicht so ernst sind wie die erste Diskussion. (4) Ja. (...)” (*Interview_4*)

“R: Muss man Gabalier kennen?” (*Interview_3*)

“Ja, vielleicht manchmal ein bisschen Hintergrundinformationen. Also ich konnte jetzt z.B. bei dem Buddhismus jetzt wusste ich nicht, was da jetzt genau stimmt und was nicht.” (*Interview_2*)

“am Ende war das ja hauptsächlich politische Kontexte. Da musste ich jetzt, also da konnte ich manche Namen nicht und wusste jetzt nicht in welche, also von Politikern in welche, also ob die jetzt eher links oder rechts einzuordnen sind auf dem politischen Spektrum” (*Interview_2*)

“Kommentarspalten irgendwo im Internet aussehen” (*Interview_2*)

“Also, ich habe jetzt angefangen diese Kontextsphäre zu lesen und der der Teil, wo es da der Thread, wo es um die Religion geht, geht ja relativ lang. Irgendwann bin ich da auch ein bisschen ausgestiegen aus der Konversation, weil da schon viele für mich fachfremde Gebiete waren” (*Interview_1*)

“dass das für mich ein bisschen fern wirkt alles und ansonsten” (*Interview_1*)

“ich bin mir nicht ganz sicher, wie die, wie die sich, also wenn jetzt auch so mit Hinblick auf politische Meinung, wie die da aufgestellt sind, aber ich glaube, die ist relativ liberal” (*Interview_1*) "ich noch ein bisschen mehr auch drauf eingehen, wie das Forum der Standard generell funktioniert, wie wie man da miteinander umgeht, wie man in Funk generell miteinander umgeht." (*Interview_1*)

#### Overwhelming

"Also, es war inhaltlich viel, einfach erstmal." (*Interview_5*)

"R: Ist schon sehr viel." (*Interview_4*)

"Ich fand es ziemlich schwer zum zum Teil mit den Kommentaren klar zu kommen. Also manche fand ich jetzt. (..) Irgendwie schwierig zu verstehen. Kürzere Kommentare aber sonst fand ich fand ich alles kurz gut. Okay." (*Interview_4*)

"irgendwie 400 Kommentare, wenn man dann jeden so analysieren würde, es ist sehr umfangreich einfach und das dann halt irgendwie in einem in einer Datensammlung zu generalisieren oder irgendwie einzuschätzen, finde ich schwierig." (*Interview_3*)

"also es ist viel zum Nachdenken so parallel, also wenn man sich so die Kommentare alle durchliest, dann versucht man so zu verstehen oder viele Informationen auf einmal auf jeden Fall. Genau, wenn man so Kontext und alles so im Kopf hat." (*Interview_2*)

"weil ich jetzt auch während der Lesezeit bestimmt vier Fünftel der Zeit mit den Kommentaren verbracht habe und deswegen ist es ein guter guter Punkt anzufangen" (*Interview_1*)

#### Coherent and Intuitive

"Für mich hat es irgendwie auch diesen Beigeschmack, dass der scheinbar extrem viel auf diesen Foren unterwegs ist, auch wenn man sich anguckt. Also es sind ja schon sehr diese Diskussionen, die die führen, die das ist ja schon sehr schneller Schlagabtausch, sage ich mal. I: Ja. Woran erkennst du das? R: Ja, an den Zeiten halt." (*Interview_3*)

"bin ich mal rüber gesprungen zur Analyse. Und da ist mir auch aufgefallen, dass das Sinn macht auf jeden Fall, das so ein bisschen zu unterteilen und sich auch die Data Source erstmal anzuschauen", (*Interview_1*)

"Wo waren das? Das stand, glaube ich in der Analyse auch." (*Interview_1*)

"Generell finde ich es gut auf jeden Fall, wie sie strukturiert ist, dass die Themen eingeordnet werden" (*Interview_1*)

### 7.8.3 C: Utility for Content Moderation

#### Not Useful for Moderation

"R: Ich wüsste jetzt nicht, was es halt für ein Mehrwert für die Moderation hätte. R: Also wenn du vor allem den Leuten halt nicht irgendwas verbieten möchtest." (*Interview_3*)

### Useful for Moderation

"Ja. Also, diese Kommentare zu sich so durchzulesen und die diese Verläufe ist halt super zeitaufwendig und das dann zu abstrahieren, in was da eigentlich inhaltlich drin steckt, wäre glaube ich anstrengend. Also, ich würde sagen, diese Analyse nutzen und dann aber wenn da was Auffälliges ist, halt in die Kommentare reinschauen, um zu verstehen, was bei das Problem ist, und inwiefern man da eingreifen muss." (*Interview_5*)

"Ich würde natürlich vielleicht anfangs beides zu Hand nehmen und schauen, ob es ob es zutrifft. Aber ich so ich finde jetzt die die Analyse stimmt schon wirklich gut und darum würde ich schlussendlich nur die Analyse verwenden. I: Das würde es noch mal kontrollieren." (*Interview_4*)

"die wäre quasi zu lang im Kontext zum Beispiel Facebook oder sowas, als dass du sie dann für jeden Nutzer, die reingucken würdest. R: Ja einfach nur, weil es da wahrscheinlich so viele. (..) Ja, dass da das vielleicht eher nicht. (..) Nicht wirklich möglich ist. Zum dass für jeden Nutzer durchzulesen oder halt zum. Weil es so viele" (*Interview_4*)

"Na dann würde ich wahrscheinlich schon sowas vorziehen. So eine Analyse." (*Interview_4*)

"Okay, Und wenn es denn wirklich vielleicht eine extreme Analyse ist das dass es. (..) Wenn es denn so grenzwertig ist, dass die Person vielleicht nicht gut für das Forum ist, dann würde ich die Kommentare verwenden. Also nicht nur die Analyse. (...)" (*Interview_4*)

"Ich glaube, ich würde mir zuerst die allgemeine Analyse durchlesen, um so einen Überblick zu kriegen, okay, das ist so die Tonalität von User X und oder vom Zielnutzer, um dann bestimmte Aussagen besser auch einschätzen zu können, wenn man weiß, okay, so macht er das oder sie das immer. Dann könnte ich mir vorstellen, dass man bestimmte Aussagen oder Kommentare schneller einordnen kann. Und schneller weiß, okay, der das ist halt seine Art und zack. Also, ich glaube, die würde auf jeden Fall helfen, um schneller da moderieren zu können und die Person schneller zu verstehen, in welche Richtung sie vielleicht auch gehen und wie das Verhalten ist oder vielleicht auch um so antizipieren zu können. Okay, da kommt jetzt so eine Art von Kommentar, wie würde die Person darauf eingehen und so. Genau, vielleicht wäre es noch hilfreich, wenn wenn noch mehr Beispiele in der Analyse genannt würden oder vielleicht zu einem Punkt auch zwei oder drei Beispiele sind, dass man das da noch ein bisschen besseres Gefühl für kriegt, wenn man sich jetzt die Kommentare selbst nicht durchlesen würde." (*Interview_2*)

"wenn man jetzt das Ziel hat, den Nutzer halt rauszufinden, wie der sich in dem Forum der Standard verhält, um ihn z.B. zu moderieren oder gegebenenfalls auszuschließen, falls das nicht okay ist, was der schreibt, dann finde ich die Analyse schon recht gut", (*Interview_1*) “spart auf jeden Fall Zeit”, (*Interview_1*)

“weil ich jetzt auch während der Lesezeit bestimmt vier Fünftel der Zeit mit den Kommentaren verbracht habe und deswegen ist es ein guter guter Punkt anzufangen” (*Interview_1*)

“vielleicht dann in der Analyse daran denken sollte irgendwie die Beispiele noch mal explizit anzuführen und nicht nur einzelne Worte zu zitieren, so Spinner z.B.” (*Interview_1*)

### 7.8.4 D: Perceived Limitations

#### Suggested Improvments

“das liegt einfach auch an meinem Background, dann gehe ich eher davon aus, dass man halt guckt, keine Ahnung, wie viel Ärger, wie viel Freude, wie viel Überraschung, Ekel, Wehmut ist da drin und wie sehr und womit hängt das zusammen und wie ist die Interaktion mit anderen Usern.” (*Interview_5*)

“Ich hätte nicht so dieses sehr technische erwartet, von, also generell hätte ich glaube ich mehr Emotions-Begriffe einfach erwartet in der Analyse.” (*Interview_5*)

“Vielleicht irgendwie so ein bisschen noch mal auf dieses auf dieses Themenübergreifend einzugehen , zu sagen irgendwie egal, ob es jetzt um Buddhismus geht, oder um die FPÖ, oder um äh keine Ahnung, was haben wir denn da noch? Andreas Gabalier oder so, die Person schreibt immer mit keine Ahnung, so und so viel Frustration und so und so viel oder im Durchschnitt, keine Ahnung, über Kontexte hinweg ist die die Emotionen primär bei der Person oder so.”(*Interview_5*)

“ich ich bräuchte trotzdem noch mal vielleicht so das am Anfang direkt klarer, oder, wenn man das wirklich anwenden wollen würde, dann bräuchte ich, glaube ich, so ein irgendwas, was man sofort erkennt, so, das ist unproblematisch, das ist eher neutral, das ist super intensiv, da geht irgendwie heiß her in den Kommentaren, sowas, also, dass man hier ja, dass man direkt erkennt.” (*Interview_5*)

“Mhm. Ja. Ähm, aber trotzdem die Möglichkeit, das halt dann noch mal detailliert sich anzuschauen.” (*Interview_5*)

“diese menschliche Seite irgendwie mehr zu verstehen, glaube ich, und und das nicht nur so, jetzt ist es halt, glaube ich, wirklich mega abstrahiert auf so eine so eine Ebene, die, glaube ich, einfach so ein bisschen vom vom Individuum wegkommt, aber wie gesagt, das muss, das ist einfach mein, das liegt, liegt an meinen Interessen und an dem, wie ich arbeite und das ist halt sehr mit mit einem Individuum und eher zu schauen so” (*Interview_5*)

“Vielleicht dazu noch so eine generelle Einschätzung am Anfang haben. Aber sonst? I: Also was? Was wäre die generelle Einschätzung? Also, was hat dir da gefehlt? (..) R: Was? (4) So in Wien in so ganz am Anfang vielleicht so zwei, in zwei, zwei, drei Sätzen. Ganz kurz beschrieben wir was so die Grund was ich mir vorstellen kann für der für der ganzen. (..) Analyse aber sonst? (..)” (*Interview_4*)

“Ich glaube nicht, dass man jetzt bezogen auf die Angabe was besser machen könnte. Aber halt. (4) So generell Grundwissen zu den zu den Artikeln. Was? Was da so geschrieben? Ja, ich glaube, das hat jetzt so einen wirklichen Einfluss für die Emotionsanalyse. Aber jetzt?” (*Interview_4*)

“Oder ob die überhaupt so fähig sind, wie diese Person jetzt z.B. zu Sarkasmus oder ob da jetzt einfach nur ja, politische Propaganda rezitiert wird oder Das heißt, der Einsatz wäre eher auf einer generellen Ebene bezogen, sich überhaupt mal Überblick zu verschaffen.” (*Interview_3*)

“Wertesystem da vielleicht so ein Schieberegler, wo die Leute sich befinden” (*Interview_3*)

“wenn noch mehr Beispiele in der Analyse genannt würden oder vielleicht zu einem Punkt auch zwei oder drei Beispiele sind, dass man das da noch ein bisschen besseres Gefühl für kriegt, wenn man sich jetzt die Kommentare selbst nicht durchlesen würde.” (*Interview_2*)

“politische Art Radikalisierung oder irgendwie sowas feststellen kann, wenn man ein Zielnutzer oder eine Zielnutzerin halt so konstant analysiert, sage ich mal, ob sich da was in dem Verhalten ändert über die Zeit.” (*Interview_2*)

“also wie sich das über die Zeit verhält und ob es da Änderungen gibt und dann die verschiedenen Analysen sozusagen miteinander vergleicht.” (*Interview_2*)

“medizinischen psychologischen Kontext so für für Therapiegespräche, um zu gucken, okay, mit welchem Affektivität hat man es hier zu tun, sage ich mal, wie ist das Gegenüber, dass man sich selbst vielleicht auch darauf einstellen kann, wie man, wie man damit umgeht und wie man darauf antwortet, sage ich mal. Voll, könnte ich mir voll vorstellen” (*Interview_2*)

“analysieren und dann halt auch noch mal irgendwie eine Ebene tiefer gehen und ja, beobachten, wie wenn man die Nutzer dann irgendwie kategorisiert, wie zwei Kategorien miteinander interagieren und was das für Auswirkungen hat.” (*Interview_1*)

### Limitations of Analysis

“weil auch einfach die Überschrift emotional analysis ist und ich dann so ein bisschen mehr also, die Emotionalität mir so relativ schwach ausgeprägt vorkam oder ich finde einfach, also Sarkasmus, weiß auch nicht, ob ich das als Emotion bezeichnen würde.” (*Interview_5*)

“Mhm. Aber nicht in welcher Stärke, weißt du, oder wie aufgeladen das einfach, wie emotional involviert die Person ist in ihren Kommentaren.” (*Interview_5*)

“erade, oder ich habe jetzt gerade übertrieben, oder eigentlich ist es unangemessen, dann wäre diese Reflexionsfähigkeit oder die Flexibilität hätte das irgendwie einen emotionalen Bezug, aber hier ist es für mich eher echt so ein bisschen so eine Sache von Kognition und Reife oder sowas.” (*Interview_5*)

“Also mehr so ganz klare Emotionen, die jetzt nicht so Humor, Sarkasmus, das sind schon so so komplexe, auch eher kognitive Sachen, finde ich und haben so sekundär dann was mit wirklich fühlen zu tun. Also, ich würde mich eher fragen, keine Ahnung, wie ärgerlich ist die Person, wie ärgerlich ist die Person auf, oder frustriert von Politik, oder ähm wie sehr beleidigt die Person andere Leute oder sowas? Ähm, was steckt da so dahinter?” (*Interview_5*)

“Da stehen halt allgemeine Sachen, mir fehlt es etwas an spezifischen emotionalen. Wenn es eine Ertragsanalyse ist. Und nicht nur wie konstruiert jemand Sätze und Argumentationen.” (*Interview_5*)

“also wie gesagt, grundsätzlich stimmte ich grundsätzlich damit überein, was da steht, aber halt eher wirklich, wenn man so, wie in der Deutschanalyse, wenn man sozusagen wirklich die Sprache so auseinander nimmt, aber aus einer psychologischen Sicht, sozusagen, könnte ich daraus jetzt nicht viel mitnehmen.” (*Interview_5*)

“Nee, ich weiß, ich glaube nicht, dass das jetzt in dem Rahmen der Analyse umsetzbar wäre und auch das ist halt geprägt von sagen meinen fachlichen Sachen.” (*Interview_5*)

“Das heißt, du würdest dir schon auch eine verkürzte Ja, ich finde es halt, wenn du wirklich jeden User so betrachten würdest, ist es einfach zu aufwendig, sich das pro User so ein so ein Report erstmal anzulegen, klar, aber wenn das dann irgendwann automatisiert werden könnte, selbst sich die dann durchzulesen alle, ist noch zu generell.” (*Interview_3*)

“, aber so um generelles Bild von den Foren Usern zu generieren ist, glaube ich, ein bisschen schwierig.” (*Interview_3*)

“User vielleicht besser einschätzen kann so in so einem groben Spektrum, aber Grenzen natürlich auch, dass es jetzt nicht so Detailgetreu ist, als würde man die Kommentare alle selber lesen.” (*Interview_2*)

“dass es so Einzelfälle gibt, also ich weiß nicht, ob das dann die in der Analyse auch aufploppen würde, aber dass es Einzelfälle gibt, wo es vielleicht sich anders verhält” (*Interview_1*)

“ich weiß jetzt nicht, ob das Datenschutzmäßig irgendwie so in Ordnung geht” (*Interview_1*) "im Detail dann oder im Einzelfall, glaube ich, dass dann so das menschliche, die menschliche Empathie oder ein also oder Verständnis auf jeden Fall genauer ist." (*Interview_1*)

"dann die gewonnenen Daten weiter zu verarbeiten. Also ja, müsste man dann auch überlegen so, was man mit den gewonnenen Daten macht", (*Interview_1*)

"daran denken sollte irgendwie die Beispiele noch mal explizit anzuführen und nicht nur einzelne Worte zu zitieren, so Spinner z.B.. Genau, aber ich würde auf jeden Fall damit arbeiten." (*Interview_1*)

"Ich könnte mir vorstellen, dass es vielleicht auch noch Thread abhängig ist oder je nachdem welcher Autor welchen Artikel postet, dass die Kommentare halt auch so auf den Artikel, vor allem die ersten Kommentare unter dem Artikel wahrscheinlich viel Einfluss oder dass der Artikel viel Einfluss auf die Kommentare nimmt." (*Interview_1*)

"hey, schau mal in dem in dem Kommentar war der sehr kompromissbereit und dem in dem anderen nicht. Das hat er jetzt nicht gemacht" (*Interview_1*)

"Grenzen wäre z.B. vor allem haben wir auch drüber geredet, dass er viel mit Quellen arbeitet und ich konnte, ich hatte jetzt keinen Zugriff z.B. auf auf die Quellen direkt. Ich habe es jetzt auch nicht probiert, ne? Und ich bin mir jetzt auch relativ sicher, dass die, dass das Modell jetzt auch nicht gecheckt hat, ob das, was dir da postet, vertrauenswürdig ist und ob das da wirklich drin steht oder nicht. I: Meinst du, dass ob das wirklich der echte Kommentar war oder? R: Nee, wenn er jetzt sagt hier, ich bringe das Argument und hier ist meine Quelle A, dass das Modell dann die Quelle A checkt und die Quelle auch einordnet. Ist das jetzt irgendeine Telegram Quelle oder ist das ein wissenschaftliches Paper oder ist das von Kurier oder hat er da gerade irgendwas auf Paint gemalt und ein Screenshot davon gemacht so." (*Interview_1*)

### 7.8.5 E: Comments and Suggested Applications

**Comments and Suggest Applications**

"Ja, so Reden oder sowas, so so politische Reden oder keine Ahnung, Reden überall, kannst ja auf dem auf den Filmfestspielen, oder auf einer Trauerfeier, oder auf dem Geburtstag oder so. Ähm, ja, könnte man das da machen oder in Interviewsituationen. mhm, Was noch? Hmm. Also, muss ja irgendwas Sprachliches sein, wo Ich weiß nicht, trifft das schon das was du so im Kopf hast?" (*Interview_5*)

"Also, könnte man ja auch alles sprachliche wahrscheinlich anwenden, könnte man auch in einem Therapiekontext halt machen zu sagen, so, wir haben fünf Therapiesitzungen aufgezeichnet, irgendwie fünf erste Gespräche und gucken uns an, was so ein bisschen da die die Grund Grunddynamik der Person ist. Ähm", (*Interview_5*)

“So ein bisschen rauszuzoomen, also, das was man vielleicht in der Interaktion nicht fähig ist zu erkennen, weil man da halt mit drin steckt, sondern das irgendwie aus einer anderen Perspektive zu betrachten und wirklich zu extrahieren, was über die Situation hinweg, da erkennbar ist und zwar halt so systematisch.” (*Interview_5*)

“Ja, ich könnte mir vorstellen, dass es so im Corporate Leben vielleicht gar nicht schlecht ist, wenn man so E-Mailverkehr so analysieren könnte, dass man dann irgendwie bei seinen ich weiß nicht, jetzt in meiner Branche z.B., wenn du E-Mailverkehr mit dem Bauherrn hast, um zu erkennen, wie seine emotionale Lage sind. I: Okay. R: Also ob der dann mal einen schlechten Tag hat, würde man, glaube ich, ganz gut erkennen an so E-Mails oder ob welche Themen ihn wirklich stören z.B. oder ob es halt ja, wenn es dann mal zu einem Streit in einem E-Mailverkehr kommt, ob es der, ob der Grund emotional basiert ist oder halt wirklich faktisch basiert ist. R: Das könnte, glaube ich, eine gute Anwendung sein für sowas.” (*Interview_3*)

“Hast du die jetzt händisch gemacht oder war das auf einem Modell basiert?” (*Interview_3*)

“Wenn das jetzt mich würde mal interessieren, wie das dann für andere User aussieht, also die z.B. nicht diese Kapazitäten von Reflektion haben oder nicht dieses breite Spektrum bedienen, sondern vielleicht wirklich nur auf politische oder nur auf FPÖ Kommentare eingehen oder nur auf linkspolitische Kommentare eingehen, wie da dann halt die die Analyse der Analyse Output ist, ob der dann so breitgreifend ist oder ob das dann wirklich auch vielleicht in der in der Schublade gesteckt wird.” (*Interview_3*)

“ch hatte jetzt keinen Zugriff z.B. auf auf die Quellen direkt. Ich habe es jetzt auch nicht probiert, ne? Und ich bin mir jetzt auch relativ sicher, dass die, dass das Modell jetzt auch nicht gecheckt hat, ob das, was dir da postet, vertrauenswürdig ist und ob das da wirklich drin steht oder nicht.” (*Interview_1*)

“Also man könnte z.B. auch Foren miteinander vergleichen und das Modell auf anderen Foren laufen lassen und dann halt vielleicht auch dadurch, dass man dann viele User analysiert irgendwie die politische Richtung des Forums rauskriegt oder die Foren, wie gesagt, miteinander vergleicht auch wie wie hitzig die Debatten sind. Und genau, dass man dann halt auch probiert irgendwie ja verschiedene Websites mit miteinander zu vergleichen wäre z.B. noch finde ich eine sinnvolle Anwendung.” (*Interview_1*)

# Overview of Generative AI Tools Used

Throughout this research, several Large Language Models (LLMs) were leveraged to enhance various aspects of this work. Specifically, Anthropic's Claude Sonnet 3.5 and Claude Haiku 3.5; Open AI's GPT-4o, o1, o1-mini, o3, and o3-mini; and Google's Gemini-1.5-Flash, Gemini-1.5-Pro, Gemini-2.0-Flash, Gemini-2.0-Pro-Experimental, and Gemini-2.0-Flash-Thinking-Experimental were utilized. These models were all accessed through their corresponding developer playground: Open AI Playground, Google Cloud Platform Vertex AI and Anthropic Console Workbench. All of these models were employed to test different components of our pipeline, for example, to verify the consistent performance of the Role-Play prompts across various LLMs. The reasoning model o1 by Open AI was used very similar as described by Handa et al., saying that 57% of LLM usage attributes to augment the human capabilities for learning and iterations on output [HTM+25]. As a business informatics student lacking formal psychology training, essential domain-specific knowledge can be provided by the LLM, enabling us to understand psychological concepts like the Theory of Constructed Emotion, ask informed questions, receive consistent feedback, and effectively implement this novel solution. All of the mentioned LLMs were utilized at various stages, for evaluating results, validating research questions, improving the writing quality and as personal research assistant [BKK+23].

# Übersicht verwendeter Hilfsmittel

Während dieser Forschung wurden mehrere Large Language Models (LLMs) genutzt, um verschiedene Aspekte dieser Arbeit zu verbessern. Insbesondere wurden Claude Sonnet 3.5 und Claude Haiku 3.5 von Anthropic, GPT-4o, o1, o1-mini, o3 und o3-mini von Open AI sowie Gemini-1.5-Flash, Gemini-1.5-Pro, Gemini-2.0-Flash, Gemini-2.0-Pro-Experimental und Gemini-2.0-Flash-Thinking-Experimental von Google verwendet. Der Zugriff auf diese Modelle erfolgte über den jeweiligen Developer Playground: Open AI Playground, Google Cloud Platform Vertex AI und Anthropic Console Workbench. Alle diese Modelle wurden verwendet, um verschiedene Komponenten unserer Pipeline zu testen, zum Beispiel um die Leistung der Role-Play prompts über verschiedene LLMs hinweg zu überprüfen. Das Reasoning-Modell o1 von Open AI wurde ähnlich wie von Handa et al. beschrieben verwendet, die zeigen konnten, dass 57% der nutzung LLM auf das Lernen und zur Iteration des LLM Outputs zurück zu führen ist. Als Student im Bereich Wirtschaftsinformatik ohne Psychologieausbildung konnte uns das LLM domänenspezifisches Wissen vermitteln, das uns in die Lage versetzt, psychologische Konzepte wie die Theory of Constructed Emotion zu verstehen, fundierte Fragen zu stellen, konsistentes Feedback zu erhalten und diese neuartige Lösung effektiv umzusetzen. Alle genannten LLMs wurden in verschiedenen Phasen eingesetzt, um Ergebnisse zu bewerten, Forschungsfragen zu validieren, die Schreibqualität zu verbessern und als persönlichen Forschungsassistenten.

# List of Figures

# List of Tables

# Bibliography


[AM20] Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987, August 2020.

[AMU17] Muhammad Abdul-Mageed and Lyle Ungar. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[Ant24] Anthropic. Use XML tags to structure your prompts, 2024.

[AS07] Saima Aman and Stan Szpakowicz. Identifying Expressions of Emotion in Text. In Václav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg, 2007. Springer.

[ATT12] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions. *Science*, 338(6111):1225–1229, November 2012. Publisher: American Association for the Advancement of Science.

[Bar06a] Lisa Feldman Barrett. Are Emotions Natural Kinds? *Perspectives on Psychological Science*, 1(1):28–58, March 2006. Publisher: SAGE Publications Inc.

[Bar06b] Lisa Feldman Barrett. Solving the Emotion Paradox: Categorization and the Experience of Emotion. *Personality and Social Psychology Review*, 10(1):20–46, February 2006.

[Bar12] Lisa Feldman Barrett. Emotions are real. *Emotion*, 12(3):413–429, 2012. Place: US Publisher: American Psychological Association.

[Bar17a] Lisa Feldman Barrett. *How Emotions Are Made: The Secret Life of the Brain*. Pan Macmillan, March 2017. Google-Books-ID: vjIvDQAAQBAJ.

[Bar17b] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017. Publisher: Oxford University Press.

[Bar22] Lisa Feldman Barrett. Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist*, 77(8):894–920, November 2022.

[BCCNEA20] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, October 2020. arXiv:2010.12421 [cs].

[BHS24] Elijah Berberette, Jack Hutchins, and Amir Sadovnik. Redefining "Hallucination" in LLMs: Towards a psychology-informed framework for mitigating misinformation, February 2024. arXiv:2402.01769 [cs] version: 1.

[BK18] Svend Brinkmann and Steinar Kvale. Doing interviews. 2018. Publisher: SAGE Publications Ltd.

[BKK+23] Bastian Burger, Dominik K. Kanbach, Sascha Kraus, Matthias Breier, and Vincenzo Corvello. On the use of AI-based tools like ChatGPT to support management research. *European Journal of Innovation Management*, 26(7):233–241, December 2023.

[BLG07] Lisa Feldman Barrett, Kristen A. Lindquist, and Maria Gendron. Language as context for the perception of emotion. *Trends in Cognitive Sciences*, 11(8):327–332, August 2007. Publisher: Elsevier.

[BMG11] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in Emotion Perception. *Current Directions in Psychological Science*, 20(5):286–290, October 2011. Number: 5 Publisher: SAGE Publications Inc.

[Bry19] Marc Brysbaert. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109:104047, December 2019.

[BW21] Lisa Feldman Barrett and Christiana Westlin. Navigating the science of emotion. In *Emotion measurement*, pages 39–84. Elsevier, 2021.

[Cra21] Kate Crawford. The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence, 2021.

[CWW+24] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45, March 2024.

[CZ10] Hsinchun Chen and David Zimbra. AI and Opinion Mining. *IEEE Intelligent Systems*, 25(3):74–80, May 2010. Conference Name: IEEE Intelligent Systems.

[CZS+24] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, March 2024. arXiv:2403.04132 [cs].

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[DLD+24] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A Survey on In-context Learning, October 2024. arXiv:2301.00234 [cs].

[EA02] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128(2):203–235, 2002. Place: US Publisher: American Psychological Association.

[EF69] Paul Ekman and Wallace V. Friesen. The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica*, 1(1):49–98, January 1969. Publisher: De Gruyter Mouton.

[EF71] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. Place: US Publisher: American Psychological Association.

[Eis89] Robert Eisenberger. *Blue Monday: The Loss of the Work Ethic in America.* IICA, 1989. Google-Books-ID: dM5YaaHzSw4C.

[Ekm92] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, May 1992.

[Ekm04] Paul Ekman. Emotions revealed. *BMJ*, 328(Suppl S5):0405184, May 2004. Publisher: British Medical Journal Publishing Group Section: Student.

[FNJL24] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as You Desire. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[GHR+24] Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. LLM-based NLG Evaluation: Current Status and Challenges, February 2024. arXiv:2402.01383 [cs].

[Gna21] Timo Gnambs. The Day of the Week Effect on Subjective Well-Being in the European Social Survey. *Zeitschrift für Psychologie*, February 2021. Publisher: Hogrefe Publishing.

[Goo24a] Google. Controlled generation | Generative AI on Vertex AI, November 2024.

[Goo24b] Google. Gemini 1.5 Pro 2M context window, code execution capabilities, and Gemma 2 are available today- Google Developers Blog, June 2024.

[Goo24c] Google. Our next-generation model: Gemini 1.5, February 2024.

[Goo25a] Google. Gemini 2.0 is now available to everyone, February 2025.

[Goo25b] Google. Long context | Generative AI on Vertex AI, January 2025.

[Goo25c] Google. Structure prompts | Generative AI on Vertex AI, January 2025.

[GRVDVB14] Maria Gendron, Debi Roberson, Jacoba Marieta Van Der Vyver, and Lisa Feldman Barrett. Cultural Relativity in Perceiving Emotion From Vocalizations. *Psychological Science*, 25(4):911–920, April 2014.

[HC10] Alan Hevner and Samir Chatterjee. Design Science Research in Information Systems. In Alan Hevner and Samir Chatterjee, editors, *Design Research in Information Systems: Theory and Practice*, pages 9–22. Springer US, Boston, MA, 2010.

[Hev07] Alan R Hevner. A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4, 2007.

[HMPR04] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, pages 75–105, 2004. Publisher: JSTOR.

[HRK+24] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X. Wang, and Sadid Hasan. Does Prompt Formatting Have Any Impact on LLM Performance?, November 2024. arXiv:2411.10541 [cs].

[HTKHJ24] Steffen Herbold, Alexander Trautsch, Zlata Kikteva, and Annette Hautli-Janisz. Large Language Models can impersonate politicians and other public figures, July 2024. arXiv:2407.12855 [cs].

[HTM+25] Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. February 2025.

[HXX+22] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner Monologue: Embodied Reasoning through Planning with Language Models, July 2022. arXiv:2207.05608 [cs].

[Iza13] Carroll E. Izard. *Human Emotions*. Springer Science & Business Media, November 2013. Google-Books-ID: DYoHCAAAQBAJ.

[JLF+23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, March 2023.

[KLH24] Terry Koo, Frederick Liu, and Luheng He. Automata-based constraints for language model decoding, August 2024. arXiv:2407.08103 [cs].

[KZC+24] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better Zero-Shot Reasoning with Role-Play Prompting, March 2024. arXiv:2308.07702 [cs].

[Lan24] Langsmith. LangSmith, September 2024.

[LB08] Kristen A. Lindquist and Lisa Feldman Barrett. Constructing Emotion: The Experience of Fear as a Conceptual Act. *Psychological Science*, 19(9):898–903, September 2008. Publisher: SAGE Publications Inc.

[Lin04] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[LJH+25] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge, January 2025. arXiv:2411.16594 [cs].

[Mat90] David Matsumoto. Cultural similarities and differences in display rules. *Motivation and Emotion*, 14(3):195–214, September 1990.

[May14] Philipp Mayring. Qualitative content analysis: theoretical foundation, basic procedures and software solution. 2014. Publisher: AUT.

[MPH+19] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825, July 2019.

[MRCZ12] A. Moreo, M. Romero, J. L. Castro, and J. M. Zurita. Lexicon-based Comments-oriented News Sentiment Analyzer system. *Expert Systems with Applications*, 39(10):9166–9180, August 2012.

[MS15] Addlight Mukwazvure and K.P Supreethi. A hybrid approach to sentiment analysis of news comments. In *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, pages 1–6, September 2015.

[MSS+24] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

[MT13] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.2012.00460.x.

[NDCR17] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why We Need New Evaluation Metrics for NLG, July 2017. arXiv:1707.06875 [cs].

[NN23] Irina Nalis and Julia Neidhardt. Not Facial Expression, nor Fingerprint – Acknowledging Complexity and Context in Emotion Research for Human-Centered Personalization and Adaptation. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, pages 325–330, New York, NY, USA, June 2023. Association for Computing Machinery.

[NN24] Friday Nyimbili and Leah Nyimbili. Types of Purposive Sampling Techniques with Their Examples and Application in Qualitative Research Studies. *British Journal of Multidisciplinary and Advanced Studies*, 5(1):90–99, February 2024.

[Ope24] OpenAi. Structured Output, 2024.

[PHMM23] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*, 14(1):108–132, January 2023. Conference Name: IEEE Transactions on Affective Computing.

[Pic00] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[Plu01] Robert Plutchik. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001. Publisher: Sigma Xi, The Scientific Research Society.

[PMH+18] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines. *IEEE Intelligent Systems*, 33(6):17–25, November 2018.

[PR00] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, December 2000. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[RB99] James A. Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999. Publisher: American Psychological Association.

[Rus80] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. Place: US Publisher: American Psychological Association.

[Rus03] James A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, 2003. Place: US Publisher: American Psychological Association.

[SCTS24] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting, July 2024. arXiv:2310.11324 [cs].

[SHQ19] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence, March 2019. arXiv:1903.09588 [cs].

[SM07] Carlo Strapparava and Rada Mihalcea. SemEval-2007 Task 14: Affective Text. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[SMR23] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-Play with Large Language Models, May 2023. arXiv:2305.16367 [cs].

[Sta21] Karen M Staller. Big enough? Sampling in qualitative inquiry. *Qualitative Social Work*, 20(4):897–904, July 2021. Publisher: SAGE Publications.

[TGL+24] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and others. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[TKC01] Y.-I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[Tom62] Silvan Tomkins. *Affect Imagery Consciousness: Volume I: The Positive Affects*. Springer Publishing Company, January 1962. Google-Books-ID: WIpgNerqaIkC.

[TWT+24] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models, October 2024. arXiv:2408.02442 [cs].

[Tö23] Petter Törnberg. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning, April 2023. arXiv:2304.06588 [cs].

[vdLGvM+19] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In Kees van Deemter, Chenghua Lin, and Hiroya Takamura, editors, *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan, October 2019. Association for Computational Linguistics.

[VSP+23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs].

[VWR24] Anton Voronov, Lena Wolf, and Max Ryabinin. Mind Your Format: Towards Consistent Evaluation of In-Context Learning Improvements, June 2024. arXiv:2401.06766 [cs].

[Wer24] Hannes Werthner. Digital Transformation, Digital Humanism: What Needs to Be Done. *Hannes Werthner · Carlo Ghezzi · Jeff Kramer · Julian Nida-Rümelin · Bashar Nuseibeh · Erich Prem ·*, page 115, 2024.

[WL23] Brandon T. Willard and Rémi Louf. Efficient Guided Generation for Large Language Models, August 2023. arXiv:2307.09702 [cs].

[WWS+22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837, December 2022.

[YSZ17] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaiane. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Comput. Surv.*, 50(2):25:1–25:33, May 2017.

[ZCS+23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023. arXiv:2306.05685 [cs].

[ZWW+24] Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. DialogueLLM: Context and Emotion Knowledge-Tuned Large Language Models for Emotion Recognition in Conversations, January 2024. arXiv:2310.11374 [cs].

[ZYY24] Yifan Zhang, Yang Yuan, and Andrew Chi-Chih Yao. Meta Prompting for AI Systems, June 2024. arXiv:2311.11482 [cs].