Harnessing Generative AI for Enhanced Sentiment Analysis in Organizational Settings

Leveraging Prompt Engineering and Adaptive Models for Smarter, Faster Workplace Insights

FROM

HRTec
Assessment System (HAS)

Presented at the 2025 Society for Industrial and Organizational Psychology Annual Conference, Denver, CO

SIOP ANNUAL CONFERENCE
DENVER, COLORADO • April 2-5, 2025

Contents

Introduction

The rapid advancement of generative AI, particularly in natural language processing (NLP), has opened up new possibilities in text generation and sentiment analysis. Industrial-Organizational practitioners are leveraging generative AI across various aspects of organizational assessment, including survey development and data analysis (Meaden, Sturdivant & Theys, 2024). Traditional comment analysis, particularly when dealing with large volumes of open-ended responses, is a time-consuming and resource-intensive process. Manual content analysis requires significant human effort to ensure accuracy and consistency in coding and interpreting responses (Vaismoradi, Turunen, & Bondas, 2013). This process, while thorough, can be prone to subjectivity and inconsistencies due to varying interpretations by analysts (Erlingsson & Brysiewicz, 2017).

Generative AI presents an opportunity to streamline this process. By automating aspects of text analysis, such as coding and theme identification, AI can significantly reduce the time and effort required from human analysts. The integration of AI in comment analysis accelerates the process and enhances consistency and objectivity (Stone, Deadrick, Lukaszewski, & Johnson, 2015). Furthermore, AI tools can continuously learn and adapt, improving their accuracy over time and reducing the likelihood of human error (Chamorro-Premuzic, Akhtar, Winsborough, & Sherman, 2017).

As generative AI continues to evolve, its potential to revolutionize traditional comment analysis becomes increasingly evident. However, despite their powerful capabilities, AI models often produce errors that can undermine their reliability and accuracy. Common errors include generating misleading or biased text, misunderstanding context, and incorrectly analyzing sentiment. As these models become more integral to organizational assessments, it's crucial to address these issues to fully harness their potential.

Prompt engineering, the process of designing and refining input prompts to guide AI models towards desired outputs (Meaden, Sturdivant, & Theys, 2024), has emerged as a potential solution to these challenges. This paper explores the nature of errors in generative AI text and sentiment analysis, examines the potential of prompt engineering to mitigate these errors, and discusses the limitations and future directions of this approach.

Purpose

The current study aims to compare the performance of ChatGPT with human classification and sentiment analysis on text data collected from an employee survey. A key focus is on prompt engineering, evaluating how tailored prompts can optimize ChatGPT's accuracy in analyzing and understanding the text data. This project includes an interdisciplinary approach, drawing on expertise from I/O psychology and from data science for more accurate and insightful sentiment analysis, making the findings both practical and data-driven.

Research Question 1: **How does ChatGPT's performance compare with human classification in sentiment analysis?**

Research Question 2: **Can "prompt engineering" optimize ChatGPT's accuracy?**

Background

Sentiment analysis can be approached in three ways: Traditional, Human-in-the-Loop, and AI-Human Learning. The Traditional method relies entirely on human reviewers to classify sentiment, ensuring accuracy but requiring significant time and effort. Human-in-the-Loop combines AI automation with human oversight to improve reliability, while Human Learning involves training AI on labeled data, allowing it to refine its performance over time.

Approaches to sentiment analysis

- Traditional: Relies entirely on humans to classify sentiment in text data.
- Human-in-the-Loop: Combines AI processing with human oversight, where subject matter experts review and correct AI classifications to improve accuracy and reliability.
- AI-Human Learning: Involves humans coding a significant portion of the data to train the AI, which is then validated on a smaller dataset to assess and refine its performance.

Large Language Models (LLMs) are trained on massive text datasets using deep neural networks to understand and generate human-like text. This training requires significant computational power, data storage, expertise, and advanced machine learning algorithms and infrastructure, often only available to tech giants like OpenAI, Google, and Microsoft. LLMs serve as foundational AI models that can be adapted for various applications, including text generation, summarization, and sentiment analysis.

For companies without the resources to build their own LLMs, pre-trained models such as OpenAI's ChatGPT provide a practical solution. **ChatGPT is an example of an LLM that has been fine-tuned for usability in conversational AI and text-based tasks, making it more accessible for businesses without extensive AI expertise.** These **pre-trained and fine-tuned** models can be integrated into workflows for tasks such as data analysis, enabling businesses to leverage AI without incurring high development costs. Some studies have shown that LLMs perform well in sentiment analysis, capturing nuanced emotions at the expense of computational efficiency (Botunac, Brkić Bakarić, & Matetić, 2024). As shown in **Table 1**, ChatGPT can classify sentiment across various expressions, accurately distinguishing between positive, negative, neutral, and mixed emotions in text.

Table 1. Example ChatGPT Sentiment Analysis.

Text data	GPT Predicts…
My awesome manager	POSITIVE
My evil manager	NEGATIVE
My adequate manager	NEUTRAL
My annoying yet effective manager	MIXED

What happens when you use AI to conduct sentiment analysis?

This table illustrates how AI classifies sentiment based on text data, highlighting its ability to recognize **positive, negative, neutral, and mixed sentiments**. While AI can accurately classify straightforward statements (e.g., *"My awesome manager"* as **positive** and *"My evil manager"* as **negative**), it faces challenges with **nuanced or mixed sentiments**, such as *"My annoying yet effective manager"*. This demonstrates the importance of **context and prompt engineering** in improving AI-driven sentiment analysis.

While general-purpose LLMs like ChatGPT offer powerful capabilities, they come with drawbacks:

1. General-purpose LLMs are trained on general datasets and may not perform optimally in specialized domains or handle industry-specific jargon.
2. Pre-trained models like ChatGPT may not have access to real-time data, meaning they can provide outdated information.[1]
3. Privacy and security concerns also arise, as sensitive or proprietary data might be processed by third-party services (Cooper et al., 2024).
4. LLMs can provide incorrect output due to inherent biases in training data, making them less reliable for critical decision-making without proper oversight.
5. LLMs also struggle with interpreting emotions from text, often missing nuances like sarcasm or ambiguity in emotional expression.
6. Integrating pre-trained LLMs into workflows can also be costly, as subscription fees may accumulate over time, especially with heavy usage.

—

Mitigating Errors with Prompt Engineering

By crafting input that guides generative AI models toward desired outcomes, prompt engineering helps reduce errors and improve the accuracy of text and sentiment analysis. Providing specific instructions or contextual cues is key to this process, ensuring more precise results. Key strategies include:

1. Using more specific prompts to help the model understand the task better. For example, “Analyze the sentiment of employee feedback regarding work-life balance” is more effective than a general prompt like “Analyze the sentiment of employee feedback.”
2. Structuring prompts to encourage logical flow, such as breaking complex tasks into smaller steps or using follow-up prompts, can address incoherent outputs.
3. Providing examples of sarcasm or irony, like “Analyze this sentence: ‘I highly recommend this place for anyone who loves being micromanaged!’ (sarcasm, negative),” enables the model to better interpret complex emotions.
4. In cases of mixed sentiment, prompts can ask the model to recognize and address ambiguity, enhancing its accuracy.
5. Providing examples of the “ground truth” (human-labeled data, called “Human Count” in our results) in a “human-learning approach”. This process involves feeding the model with a dataset to help it adapt its language understanding and responses to align with the analysis goals and requirements. The dataset is split into two sets randomly: a training set and a validation or test set. The chosen model then uses the training set to learn the patterns of the data, and its performance is evaluated on the test data. Usually, it is considered best practice to split the data such that 80 or 70% of the data comprises the training data, and 20 or 30% of the data comprises the test data (Gholamy, Kreinovich, & Kosheleva, 2018).

By applying these techniques, prompt engineering can likely improve AI-driven sentiment analysis.

[1] While generative AI models like ChatGPT have fixed knowledge bases, this limitation is not directly relevant to sentiment analysis tasks, which focus on interpreting the text's emotional tone rather than requiring real-time updates. Thus, our study emphasizes the model's ability to classify sentiment based on linguistic features, regardless of its static knowledge.

Current Study

When approaching sentiment analysis for survey responses, choosing the right model involves weighing factors like data availability, complexity, interpretability, and resource constraints. Given our organizational constraints, our access to text data in the context of employee experience surveys, and the nature of the task, we determined that using a pre-trained language model (ChatGPT) was appropriate. We conducted this project in two phases: first, an exploratory study to observe how well ChatGPT analyzes sentiment with minimal instructions, and second, a comparison of human coding results with ChatGPT's performance when using detailed prompt engineering. This study is exploratory in nature, designed to investigate the potential and limitations of generative AI, particularly ChatGPT, in performing sentiment analysis. As shown in **Table 2**, our approach integrates both I/O psychology and data science methodologies, combining qualitative insights with computational techniques to enhance AI-driven sentiment analysis. By leveraging a unique dataset from an I/O psychology context, we aim to better understand how general-purpose language models perform on nuanced text data and to identify areas where prompt engineering and hybrid AI-human methods may improve outcomes.

Table 2. Interdisciplinary Approach of the Current Study.

		I/O Psychology	Data Science
Background	Qualitative Insights	✓	
	Quantitative Rigor	✓	✓
Methods	Prompt Engineering		✓
	Sentiment Analysis	✓	✓
Implications	Understanding Employee Sentiment	✓	
	Data-Driven Decisions	✓	✓
	Application of AI	✓	✓
	Optimizing AI		✓

This study combines insights from **Industrial-Organizational (I/O) psychology** and **Data Science** to enhance AI-driven sentiment analysis. I/O psychology provides expertise in survey design and human interpretation of workplace sentiment, while Data Science leverages AI models to process large-scale text data efficiently. By integrating these disciplines, we refine AI's ability to classify sentiment with greater accuracy and practical relevance for organizational decision-making.

Phase 1: Human-in-the-Loop Baseline

We began with an exploratory analysis to evaluate ChatGPT's sentiment analysis capabilities using two straightforward prompts. The goal was to assess how effectively ChatGPT could classify sentiment in open-ended text responses without relying on extensive prompt engineering or customization. The insights from this phase will inform the more detailed prompt engineering approaches applied in the second phase of this research.

—

Method

Sample

Our approach focuses on exploring ChatGPT's performance in a four-class sentiment analysis task using iterative prompt engineering. The goal was not to achieve optimal accuracy but rather to evaluate the model's strengths, weaknesses, and potential applications in organizational research.

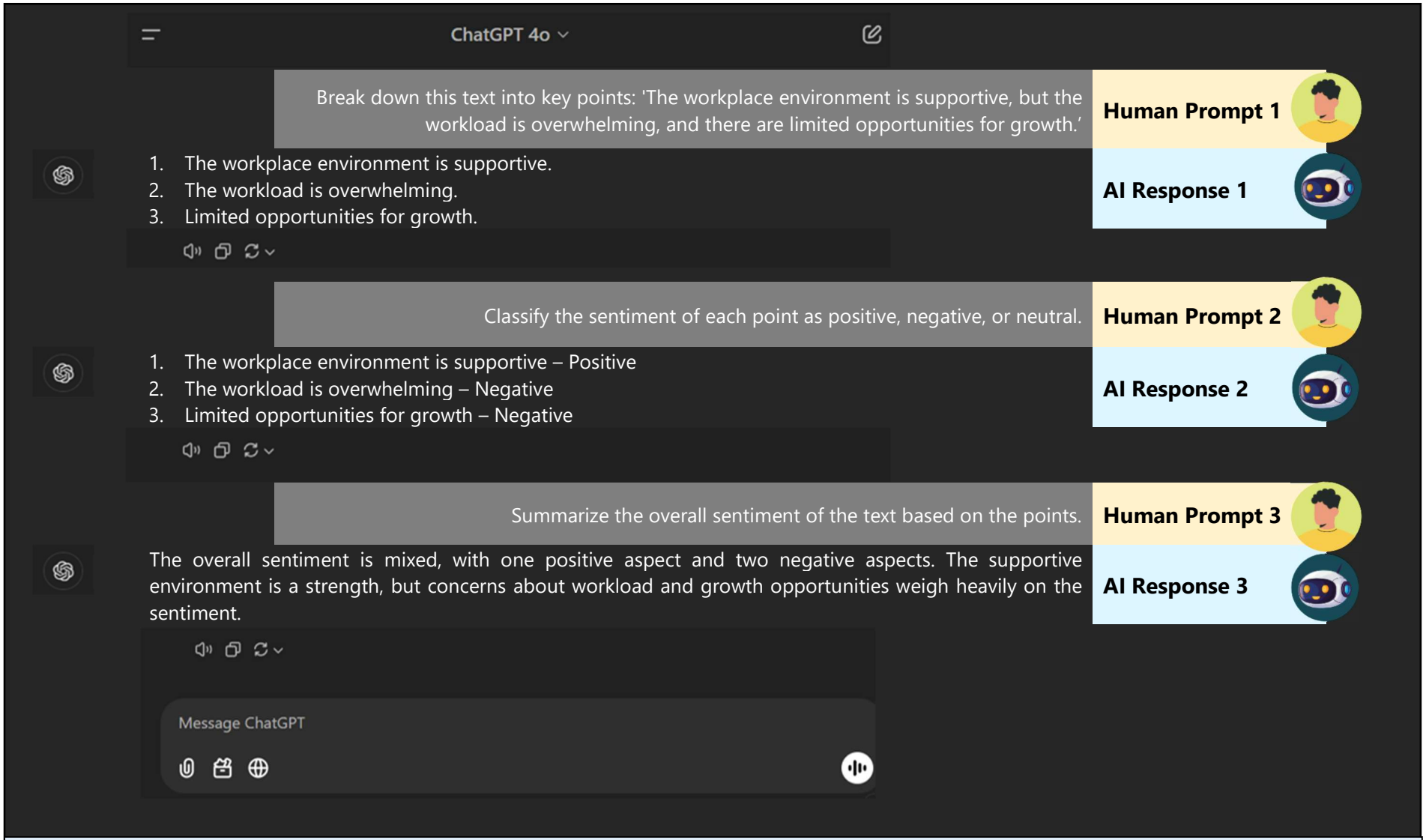
Employees from two regional offices of a real-world organization, who are typically relocated to the region on a temporary basis for 1-2 years, responded to an open-ended question in an employee experience survey. The question asked employees to describe their experiences in relation to the organization's core values during their relocation, and whether any changes were needed in the application of these values within their current work environment. This resulted in 138 comments.

Procedure

Our procedures included the following steps:

1. Data preparation, including extracting User ID and comments and removing organizational or personal identifiers.
2. Text data was input into ChatGPT (OpenAI's GPT-4 model via ChatGPT interface) with specific instructions which were utilized alongside "prompt chaining", or breaking down complex tasks into a series of smaller, connected prompts, where the output of one prompt serves as input for the next (OpenAI, n.d., Prompt engineering guide). This method allows ChatGPT to tackle intricate problems step by step, improving accuracy of the final output. An example of prompt chaining can be seen in **Figure 1** (the specific prompts for each iteration for Phases 1 and 2 are located in the **Appendix**). In the first prompt, ChatGPT was given context and tasked with conducting a sentiment analysis of employee survey data. In the second prompt, ChatGPT was tasked with organizing the comments and sentiment analysis into a table.
3. Employing a "human-in-the-loop" approach (Thomas, Kruse, Carson, Callahan, & Cabe, 2024), the data was exported to Excel and reviewed by an SME to identify miscategorized comments.

Figure 1. Example of prompt chain in ChatGPT.



This image demonstrates prompt chaining, a process where each user prompt builds upon the previous responses to refine or guide the AI's output. By breaking complex tasks, such as sentiment analysis of a nuanced statement, into smaller, interconnected steps, the process ensures clarity, continuity, and improved accuracy. Each prompt builds logically on the output of the previous one, making complex tasks more manageable and the results more precise.

Note: This is an example and does not reflect data or prompts used in the actual study.

Metrics

To evaluate ChatGPT's sentiment classification performance, we applied three key metrics: accuracy, a confusion matrix, and a classification table (Krstinić, Braović, Šerić, & Božić-Štulić, 2020). These metrics help assess overall correctness while also identifying where the model tends to misclassify sentiment, allowing us to evaluate ChatGPT's baseline performance and pinpoint areas where prompt engineering or human-in-the-loop adjustments may improve accuracy.

Accuracy measures the overall correctness of a model's predictions. It is calculated as the ratio of the number of correct predictions (both true positives and true negatives) to the total number of predictions (sum of true positives, true negatives, false positives, and false negatives). Essentially, accuracy measures **the percentage of correct predictions** made by the model using the following equation:

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Predictions\ (TP + TN + FP + FN)}$$

While accuracy provides a **high-level measure of performance**, it can be **misleading** in **imbalanced datasets**, where certain categories (e.g., Neutral sentiment) may dominate predictions. Therefore, we also use more **granular** performance metrics.

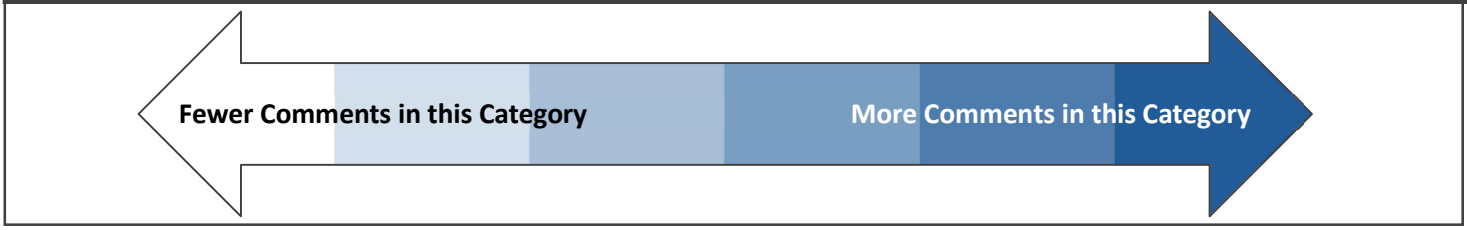
A **Confusion Matrix** is a table used in machine learning to evaluate the performance of a classification model. It organizes predictions into true positives, true negatives, false positives, and false negatives, providing insight into classification accuracy and common misclassification patterns.

Figure 2 provides a heat map color coding guide, which is useful for interpreting confusion matrices. It visually distinguishes high-frequency classifications from low-frequency ones, aiding in the analysis of ChatGPT's performance. Darker shades represent higher counts of comments classified into a specific category, while lighter shades indicate fewer instances in that category. This visual aid ensures a clearer understanding of model performance by emphasizing areas where misclassifications are frequent and where ChatGPT's sentiment predictions align most closely with human-labeled data.

Why Use a Heat Map in the Confusion Matrix?

A heat map helps quickly identify where the classification is strong and where misclassifications are frequent. It is useful for diagnosing model performance and understanding how sentiment confusion occurs. The color gradients make patterns easier to interpret at a glance compared to a plain table of numbers.

Figure 2. Heat Map Color Coding Guide.



To illustrate this further, **Table 3** presents an ideal confusion matrix, where ChatGPT achieves 100% accuracy, classifying all comments correctly with no errors. In contrast, **Table 4** provides a more realistic example, demonstrating common misclassification trends, such as ChatGPT struggling to differentiate between Neutral and Mixed sentiments or incorrectly categorizing Negative comments as Positive. Comparing these confusion matrices highlights how a perfect classification model would perform versus the real-world misclassifications observed in our study.

Table 3. Ideal Confusion Matrix: ChatGPT with 100% Accuracy.

	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	47	0	0	0
Actual Negative	0	67	0	0
Actual Neutral	0	0	19	0
Actual Mixed	0	0	0	5

Table 3. Ideal Confusion Matrix: ChatGPT with 100% Accuracy.

If ChatGPT correctly categorized all comments, this is what the confusion matrix would look like: the **confusion matrix would have nonzero values only along the diagonal**, with all off-diagonal cells containing **zeros**, indicating no misclassifications. The **heat map would display the darkest shades along the diagonal**, reflecting perfect accuracy, while the misclassification areas would remain **completely light**, signifying no errors.

Note: This table is an example and does not reflect actual project data.

Table 4. Realistic Confusion Matrix: ChatGPT with Misclassifications.

	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	48	38	61	24
Actual Negative	67	32	28	20
Actual Neutral	52	30	17	48
Actual Mixed	20	62	33	45

Table 4. Realistic Confusion Matrix: ChatGPT with Misclassifications.

Key Observations from the Example Confusion Matrix

- **Correct Predictions (Diagonal Cells):** The cells along the diagonal (e.g., 48 for Positive, 32 for Negative, 17 for Neutral, 45 for Mixed) suggest where the model is classifying correctly.
- **Misclassifications (Off-Diagonal Cells):** The 67 in Actual Negative → Predicted Positive suggests that many negative comments were misclassified as positive.
- The 62 in Actual Mixed → Predicted Negative indicates that mixed sentiments were often mistaken as negative.

Pattern Analysis

- The model struggles to differentiate between Neutral and Mixed responses, as seen in the high misclassification rates between these two categories.
- Negative comments are often misclassified as Positive, which might indicate that the model isn't effectively capturing negative sentiment cues.

Note: This table is an example and does not reflect actual project data.

A **Classification Table** is a summary table used to evaluate the performance of a classification model. It provides insights into how well the model correctly predicts each category and identifies areas for improvement. A **classification table** provides the detailed performance metrics defined in **Table 5**.

Table 5. Definitions of Reported Metrics in the Classification Table.

Metric	Precision	Recall	F1-Score
Definition (Krstinić et al., 2020)	Measures the accuracy of the positive predictions. It is the ratio of true positives to the sum of true positives and false positives. Ranging from 0 to 1, a precision of 1 means that all predicted positive instances are indeed positive, while a precision of 0 means that none of the predicted positive instances were correct.	Measures the ability of the model to capture all the relevant cases (true positives). It is the ratio of true positives to the sum of true positives and false negatives. Ranging from 0 to 1, a recall of 1 means that all actual positive instances were correctly identified, while a recall of 0 means that none of the actual positives were detected.	The harmonic mean of precision and recall, providing a single score that balances both precision and recall. Ranging from 0 to 1, an F1-Score of 1 indicates perfect precision and recall, while an F1-Score of 0 indicates the worst performance.
Formula	$\frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$	$\frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$	$2 \times \frac{Precision\ \times Recall}{Precision + Recall}$
Focus	How many of the predicted positive instances are actually positive; minimizing false positives	How many of the actual positive instances are correctly predicted; minimizing false negatives	How well a model performs overall when both false positives and false negatives matter

The classification table below compares ChatGPT's predictions (AI Count) against the Human Count. Discrepancies between these counts reveal key misclassification patterns, such as over-reliance on Neutral sentiment or difficulty in identifying Mixed sentiment. As seen in **Table 6**, the AI frequently misclassifies sentiments, leading to lower precision and recall scores, particularly in distinguishing Neutral from Mixed categories. These insights help pinpoint areas where further prompt refinement or human-in-the-loop validation may be necessary to improve classification accuracy.

Table 6. Example Classification Table.

Class	Precision	Recall	F1-Score	AI Count	Human Count
Positive	0.30	0.26	0.28	41	38
Negative	0.21	0.24	0.22	36	32
Neutral	0.14	0.11	0.12	31	32
Mixed	0.26	0.24	0.25	30	35
Overall Accuracy	—	—	0.25	—	—

Table 6. Example Classification Table.

The AI Count represents the number of times the AI predicted each sentiment category, while the Human Count reflects the actual number of comments labeled in each category by human raters. A mismatch between these values highlights misclassification trends, such as AI slightly over-predicting positive sentiment (41 AI vs. 38 human-labeled comments) while under-identifying mixed sentiment (30 AI vs. 35 human-labeled comments). The lower precision and recall scores suggest the AI struggles to differentiate between neutral and mixed comments, reinforcing the need for more refined prompt engineering or human-in-the-loop validation to improve classification accuracy.

If ChatGPT correctly categorized all comments, the **AI Count would match the Human Count exactly** for each sentiment category in the classification table, resulting in **perfect precision, recall, and F1-scores (all equal to 1.00)** across all classes.

Note: This table is an example and does not reflect actual project data.

Results

We evaluated model performance using accuracy, a confusion matrix, and a classification table, as defined in the previous section.

The SME agreed with only 26% of ChatGPT's classifications, highlighting significant misclassification errors (see **Exploratory Analysis** in the **Appendix** for a full summary). Substantial misclassifications occurred in the Neutral and Negative categories, contributing to a large number of errors.

A key finding from the confusion matrix (**Table 10**) is that ChatGPT frequently mislabeled Positive comments as Neutral (37 instances) and Negative comments as Neutral (56 instances), reinforcing its tendency to default to a Neutral classification. Additionally, Mixed sentiments were rarely detected, with nearly all Mixed comments misclassified as Negative.

As seen in **Table 11**, precision and recall scores were highly variable across categories. ChatGPT achieved high precision for Positive (0.83) and Negative (0.88) sentiments, but low recall (0.21 and 0.10, respectively), indicating that while the positive and negative predictions were often correct, it failed to identify many actual instances of these sentiments. The Neutral category had 100% recall but very low precision (0.16), suggesting that the model overclassified comments as Neutral, leading to a high rate of false positives. The Mixed category was completely misclassified (Precision = 0.00, Recall = 0.00), showing that ChatGPT was unable to recognize mixed sentiments at all.

These results indicate that ChatGPT, when using minimal prompts, struggles to classify sentiment accurately, particularly when distinguishing between Neutral and other sentiment categories. This analysis serves as a baseline for comparison in Phase 2, where we refined AI performance through prompt engineering to improve classification accuracy.

Phase 2: Prompt Engineering

This second study delved deeper into ChatGPT's sentiment analysis capabilities by incorporating extensive prompt engineering to refine and enhance its accuracy. Building on Phase 1, this study used carefully crafted prompts designed to provide more detailed instructions and examples. We compared human sentiment analysis with ChatGPT results and examined how prompt engineering may help to reduce errors.

—

Method

Sample

Employees from various locations of an organization responded to open-ended questions as part of an employee experience survey. We analyzed 292 responses to the question: "If a close friend of yours were looking for a job, would you recommend your [location] as a good place to work? Why or why not?"

Human Coding

The human coding procedure included the following (Rytting, Sorensen, Argyle, Busby, Fulda, Gubler, & Wingate, 2023):

1. Three SMEs independently coded the sentiment of each comment using the following definitions: positive for comments expressing satisfaction or positive feelings about the workplace, negative for comments expressing dissatisfaction or negative feelings, neutral for comments without a clear positive or negative sentiment, and mixed for comments containing both positive and negative sentiments.
2. The SMEs met to resolve discrepancies in their assessments. Through discussion, they refined the definitions, particularly applying a "proportion rule" to distinguish between mixed and predominantly positive or negative comments. For example, comments were classified as negative rather than mixed if a comment contained one positive remark, but the majority of the content was negative.
3. Once consensus was reached on all responses, the final dataset was compiled, ensuring a consistent and agreed-upon sentiment classification across all coders.

Prompt Engineering

This study used an iterative prompt engineering process to refine and improve output generated by ChatGPT (Leo, 2013). After sanitizing and preparing the data, text data was uploaded to ChatGPT (OpenAI's GPT-4 model via ChatGPT interface). Prompt design involved detailed descriptions of the dataset and specific instructions for handling each sentiment category, building on previous iterations and using prompt chaining as needed. See the **Appendix** for the specific prompts used.

Each time we began a new iteration, a "new chat" was used. Each iteration refined the process based on insights from the previous, beginning with a focus on handling negations (Iterations 1 and 2). Iteration 3 introduced a "proportion rule," where classification was based on the overall balance of positive, negative, and neutral tones. In Iterations 4 and 5, we used a 'human learning approach' where we applied a 70/30 split (Gholamy et al., 2018), randomly selecting 203 comments (70% of 292: 80 positive, 72 negative, 8 neutral, 43 mixed) to train ChatGPT. ChatGPT then analyzed the remaining 89 comments (30% of 292: 35 positive, 33 negative, 4 neutral, 17 mixed) using the same approach as the examples provided. We compared ChatGPT's classifications against the human-coded labels for those 89 comments. In Iterations 6 and 7, we combined the 'human learning approach' with enhanced guidelines from earlier iterations. See the **Appendix** for the results tables for each iteration.

The 70/30 split is a common guideline in machine learning for dividing data into training (70%) and testing (30%) sets, balancing learning and evaluation. While widely used, it's not a fixed rule—ratios like 80/20 or 60/40 may be better suited depending on dataset size, complexity, and task. The goal is to ensure effective training while keeping enough data for reliable evaluation.

—

Results

As in the previous phase, we used accuracy, a confusion matrix, and a classification table to evaluate the model's performance for each iteration (Krstinić et al., 2020). **Table 7** summarizes accuracy improvements across iterations. Throughout Phase 2, accuracy was improved from Phase 1. The largest gain (from 29% to 66%) occurred when integrating a human-learning approach, highlighting the value of labeled training data in refining ChatGPT's sentiment classification. The latter iterations benefited from more targeted guidance and learning from human examples, resulting in more consistent alignment with SME ratings. The model struggled to accurately classify comments as neutral or mixed (see low precision, recall, and F1-scores).

Table 7. Accuracy Metrics for Each Iteration in Phase 2.

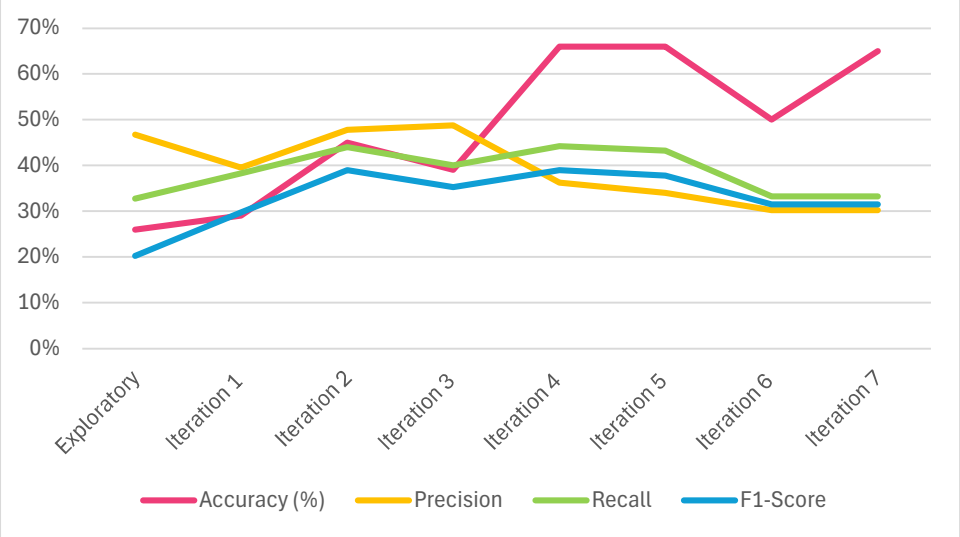
Iteration	Prompt Description	Accuracy
Exploratory	Limited Guidelines for Human-in-the-Loop Approach	26%
Iteration 1	General Guidelines	29%
Iteration 2	Handling Negations	45%
Iteration 3	Proportion Rule	39%
Iteration 4	Shift to Human Learning Approach	66%
Iteration 5	Human Learning Approach plus Guidance on Context	66%
Iteration 6	Enhanced Guidelines	50%
Iteration 7	Human Learning Approach plus Enhanced Guidelines	65%

By tailoring prompts to account for language complexity, ChatGPT demonstrated better handling of nuanced emotions and greater consistency with human coding. This study highlighted the role of prompt engineering in enhancing ChatGPT's sentiment analysis, leading to improved accuracy compared to earlier iterations, although accuracy never exceeded 66%.

Figure 3 illustrates performance trends across iterations, showing that accuracy improved significantly after incorporating human learning approaches. Initially, with minimal prompt refinement, ChatGPT achieved only 26% accuracy, indicating poor classification reliability. As prompt engineering techniques were introduced, accuracy gradually increased, with notable improvements in Iteration 2 (handling negations) and Iteration 4 (shifting to a human learning approach). The highest accuracy (66%) was observed when a human-labeled training dataset was used to fine-tune ChatGPT's classifications. However, despite improvements, the model still struggled with differentiating between neutral and mixed sentiments, as reflected in lower recall and F1-scores for these categories. These trends suggest that adding structured human input and fine-tuned prompts significantly enhanced classification accuracy, but challenges remain in distinguishing between neutral and mixed sentiments.

For complete prompts and summary of results in this phase, see sections **Iteration 1** through **Iteration 7** in the **Appendix**.

Figure 3. Performance trends across iterations.



Iteration 4 appears to have the best overall performance across the metrics.

Accuracy is high, reaching its peak after Iteration 3.

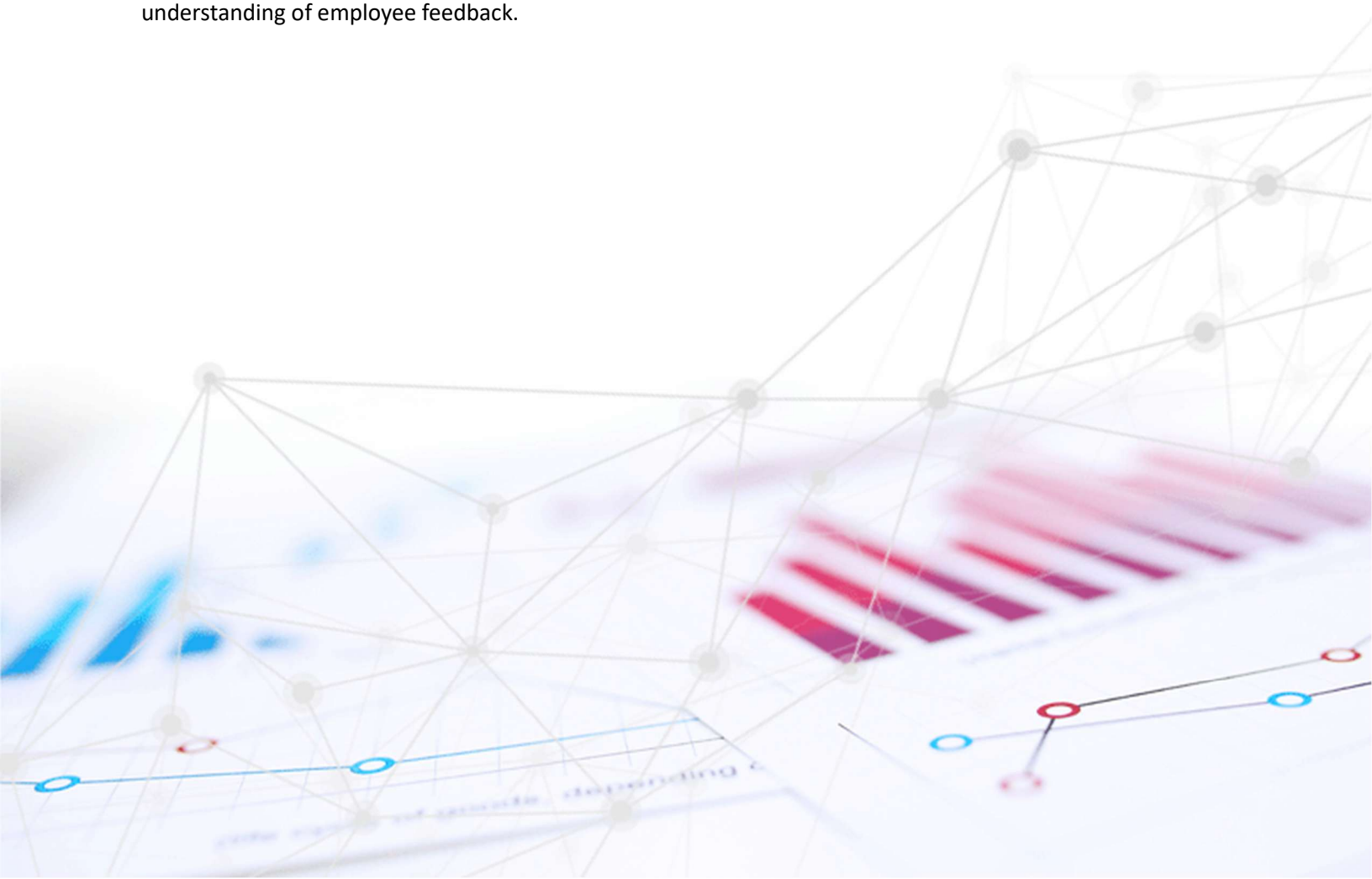
Precision, **Recall**, and **F1-Score** are closely aligned, suggesting a good balance between minimizing false positives and false negatives.

Iteration 4 reflects the point where prompt engineering optimized the model's performance before diminishing returns in later iterations.

Limitations

This study has several limitations, from both I/O psychology and data science perspectives, which may have influenced the outcomes and applicability of the findings.

1. The comments used in this study were collected from a specific organizational setting, which may not fully capture the wide range of employee experiences or sentiments found in different industries, roles, or cultural environments. Additionally, the 'human learning approach', while practical, may limit the model's performance in real-world applications where data distributions are less controlled.
2. The study relied on prompt engineering to refine the generative AI model's sentiment analysis capabilities. While this approach showed improvements, it was labor intensive and required significant human involvement to refine prompts iteratively, making the findings less scalable for organizations without access to the same level of expertise or resources. The discrepancies observed between AI-generated outputs and human coders—particularly in the classification of neutral and mixed sentiments—highlight limitations in ChatGPT's ability to accurately capture the complex emotional landscape present in employee feedback, which may impact its effectiveness for nuanced organizational assessments. Despite our efforts, accuracy never exceeded 66%, indicating persistent challenges in achieving high reliability.
3. We opted not to preprocess the data heavily—we did not lemmatize words, remove stopwords, or filter out comments that did not directly answer the research questions. This decision allowed the model to process the full context of each comment, but it may have contributed to misclassifications, especially for comments containing ambiguous statements or irrelevant information.
4. This study did not explore generative AI capabilities for thematic analysis—an essential aspect of qualitative research that involves identifying patterns or themes across data. Evaluating the potential of generative AI for thematic analysis could provide additional value, complementing sentiment analysis and offering a more holistic understanding of employee feedback.



Conclusion

Manual content analysis is often time-consuming and subject to human error, whereas AI tools offer the potential for faster, more consistent, and scalable methods to streamline traditional comment analysis. However, these models have difficulties in handling mixed sentiments, detecting negations, accurately differentiating neutral feedback, and an over-representation of positive sentiments, leading to misclassifications. As illustrated in **Table 8**, different sentiment analysis approaches—Traditional, Human-in-the-Loop, and AI Human Learning—each offer trade-offs in terms of accuracy, scalability, and automation. While additional research is needed to fully address these limitations, prompt engineering has emerged as a valuable approach to refining AI outputs and reducing errors, enhancing the model's context sensitivity, and handling emotional complexities more effectively.

Our research suggests that in addition to prompt engineering, organizations using generative AI should incorporate human review alongside automated sentiment analysis, particularly for ambiguous or off-topic responses. A hybrid approach that combines machine learning with manual coding allows for a more nuanced understanding, ensuring that mixed sentiments or sarcasm are correctly interpreted. This human-in-the-loop system helps prevent misclassification and ensures that the analysis outputs are both meaningful and actionable (Cambria et al., 2017; Medhat, Hassan, & Korashy, 2014).

Table 8. Comparing Approaches to Sentiment Analysis.

	Aspect	Traditional Sentiment Analysis	**Human-in-the-Loop Sentiment Analysis**	AI Human Learning Approach
Steps	Data Preparation	Clean and format text data	Clean and anonymize text data	Clean and anonymize text; split into 70/30 sets
	Manual Data Coding	Humans classify sentiment	None	Humans classify sentiment for 70% of data
	Model Training	None	Use pre-trained models	Train the model on manually coded data
	Input for AI Analysis	None	AI analyzes text	AI analyzes 30% of the data
	AI Sentiment Classification	None	AI classifies sentiment	AI classifies sentiment
	Human Review/Validation	Humans correct errors and refine results	Humans ensure output reliability, correct errors, validate and finalize results	Humans ensure output reliability, correct errors, validate and finalize results of initial training
	Iterative Improvement	Human analysis consists of many iterative and simultaneous processes	Refine prompts using human input	Refine or retrain models are based on human feedback
Features	Accuracy	✓ Highly accurate	✓ Higher accuracy due to human validation	− Dependent on training quality; struggles with complexity
	Cost	✗ High cost due to reliance on human involvement	✗ Higher cost due to ongoing human involvement	− High initial cost due to labor-intensive training
	Human Involvement	✗ Relies entirely on human classification	− Continuous human validation and refinement	− Human review of AI classifications for test set
	Scalability	✗ Low scalability and high labor / cost	− Moderately scalable; limited by human capacity	− Moderately scalable; limited by training effort
	Automation	✗ None	✓ Partially automated	✓ Can be set up for complete automation
Use Cases		High-stakes or complex sentiment analysis	Small-scale, context-rich analysis	Large scale, big data projects

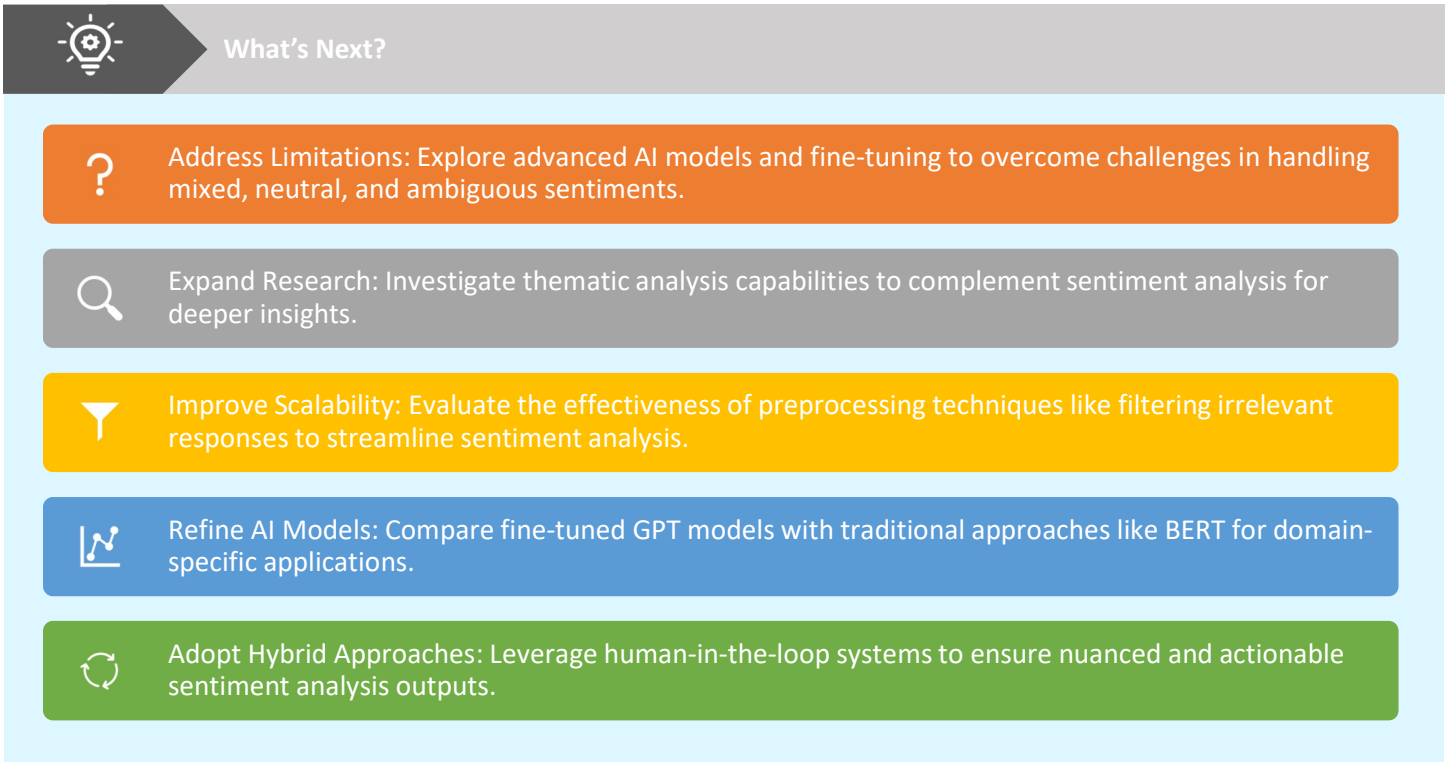
In sum, Human-in-the-Loop Sentiment Analysis balances automation efficiency with human judgment.

Future Research Directions

Future research should examine ChatGPT's advanced functions, including creating custom models known as GPTs and fine-tuning them for specific tasks. By creating a GPT, users can tailor the behavior and personality of the model to fit their unique needs, such as customer service, content generation, or technical support. Fine-tuning takes this customization further by allowing users to train the model on domain-specific data, ensuring it generates more accurate and relevant responses within a particular industry or application (OpenAI, n.d., Fine-tuning guide). Fine-tuning GPT models using domain-specific sentiment datasets, such as employee feedback data, to enhance classification accuracy and contextual understanding. Fine-tuning enables the model to adapt more effectively to the specific language patterns, tone, and unique challenges of workplace-related comments. Similar to the "human learning approach" we used in this study, this process involves feeding the model with a training dataset and a validation or test dataset (often using a 70/30 split; Gholamy et al., 2018). These functions enable more effective use of generative AI, making the models highly adaptive and personalized for specialized use cases. This could significantly improve performance, particularly in nuanced categories like mixed or neutral sentiments, where general models often struggle.

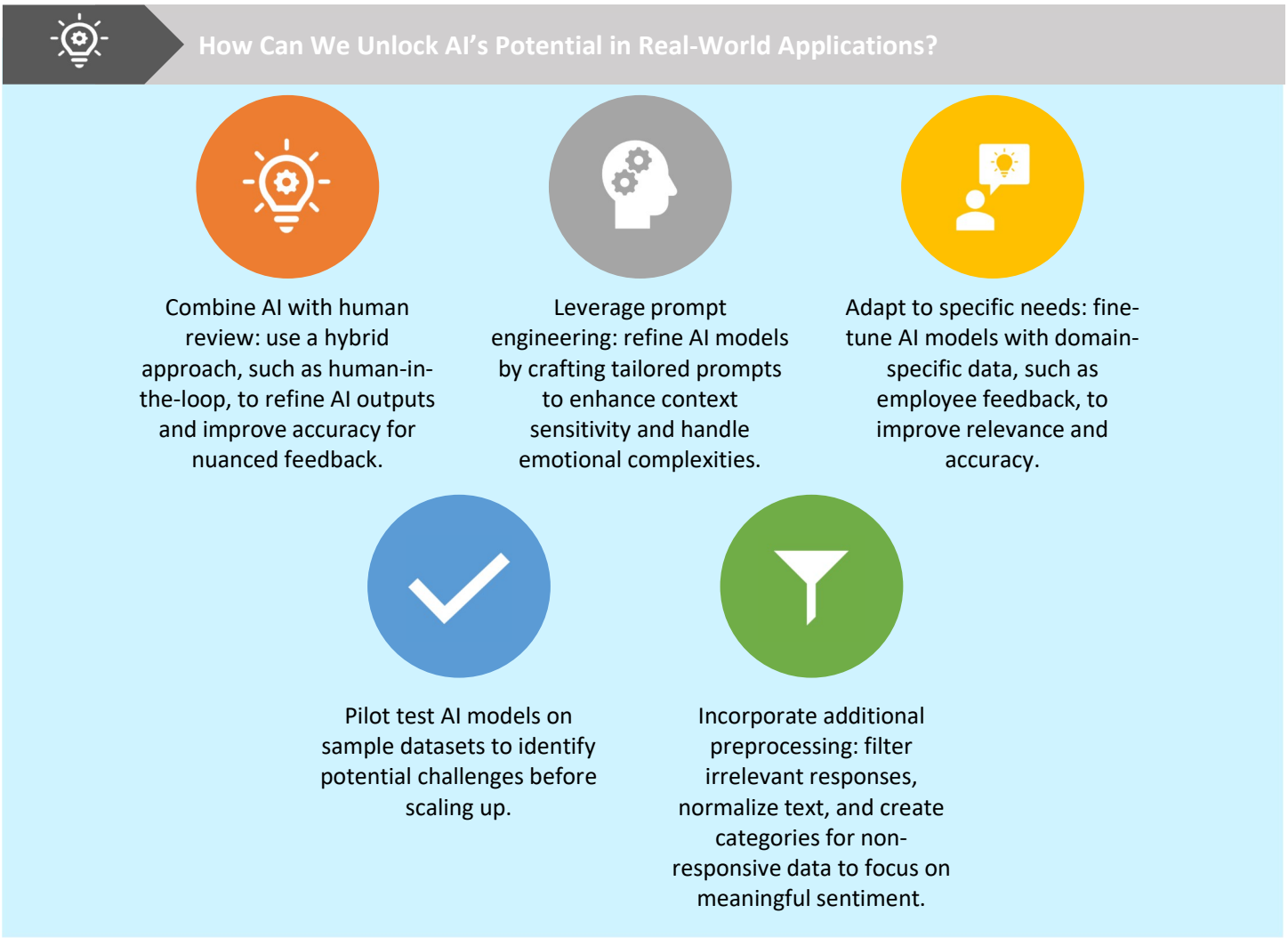
Our study briefly considered traditional pre-trained models like BERT (Bidirectional Encoder Representations from Transformers) as a point of comparison. While BERT has been effective in sentiment classification when fine-tuned on task-specific datasets (Devlin et al., 2019), generative models like GPT offer additional flexibility for tasks requiring fewer resources, such as few-shot learning or exploratory analysis. Future research could compare fine-tuned GPT models to fine-tuned BERT models to better understand their relative performance and applicability in I/O psychology contexts. Fine-tuned GPT models may achieve comparable or superior results due to their broader contextual capabilities.

Finally, research indicates that filtering out responses that lack relevant content or that don't directly address the question can enhance the quality of analysis (Liu, 2012). Therefore, we recommend that future studies evaluate the performance of generative AI sentiment analysis by preprocessing the text; this includes removing irrelevant information, identifying and handling incomplete responses, and normalizing text. Additionally, consider creating a category for responses that are neutral or irrelevant, so they can be excluded from sentiment analysis. This may allow the analysis to focus on comments where meaningful sentiment is present and prevents skewing results with non-responsive data (Pang & Lee, 2008).



Unlocking AI's Potential in Real-World Applications

Harnessing AI's capabilities for sentiment analysis and decision-making requires a balanced approach that integrates automation with human expertise. A human-in-the-loop model enhances AI's ability to interpret complex language, correct errors, and refine outputs, ensuring higher accuracy in nuanced feedback analysis. By leveraging prompt engineering, organizations can fine-tune AI models to better understand emotional complexities and context-specific nuances. Additionally, adapting AI to specific needs—such as employee sentiment analysis—through domain-specific training data improves relevance and predictive accuracy. A pilot testing approach allows organizations to identify potential limitations before full-scale implementation, ensuring that AI models are optimized for performance. Furthermore, incorporating preprocessing techniques like filtering irrelevant responses and normalizing text enhances data quality, allowing AI to focus on meaningful sentiment. By combining these strategies, businesses can unlock AI's full potential—automating large-scale sentiment analysis while maintaining human oversight for critical insights.



Final Thoughts

By leveraging prompt engineering, I/O psychologists can capitalize on AI's capabilities for sentiment analysis with minimal complexity and cost, while remaining adaptable to evolving needs. For I/O psychologists interested in using generative AI for sentiment analysis, it is recommended to start small—piloting AI models on a sample of feedback data to identify potential challenges—and to use prompt engineering to refine model outputs. By integrating structured prompt engineering and human oversight, organizations can significantly improve the reliability of AI-driven sentiment analysis, ensuring more accurate and context-aware classifications. Additionally, periodically updating models based on new organizational contexts will help ensure the accuracy and reliability of sentiment analysis, ultimately driving more effective workplace interventions. This can be a valuable asset for organizations looking to quickly gauge employee sentiment, uncover key concerns, and make data-driven decisions to improve workplace environments.

References

Beauxis-Aussalet, E., & Hardman, L. (2014, October). Visualization of confusion matrix for non-expert users. In IEEE Conference on Visual Analytics Science and Technology (VAST) - Poster Proceedings (pp. 1-2). IEEE.

Botunac, I., Brkić Bakarić, M., & Matetić, M. (2024). Comparing Fine-Tuning and Prompt Engineering for Multi-Class Classification in Hospitality Review Analysis. *Applied Sciences*, 14(14), 6254. https://doi.org/10.3390/app14146254

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15-21. https://doi.org/10.1109/MIS.2013.30

Chamorro-Premuzic, T., Akhtar, R., Winsborough, D., & Sherman, R. A. (2017). The datafication of talent: How technology is advancing the science of human potential at work. *Current Opinion in Behavioral Sciences*, 18, 13-16. https://doi.org/10.1016/j.cobeha.2017.04.005

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805

Erlingsson, C., & Brysiewicz, P. (2017). A hands-on guide to doing content analysis. *African Journal of Emergency Medicine*, 7(3), 93-99. https://doi.org/10.1016/j.afjem.2017.08.001

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *International Journal of Intelligent Technologies and Applied Statistics*, 11(2), 105-111. https://doi.org/10.6148/IJITAS.2018.1102.01

Krstinić, D., Braović, M., Šerić, L., & Božić-Štulić, D. (2020). Multi-label classifier performance evaluation with confusion matrix. *Computer Science & Information Technology*, 1, 1-14. https://doi.org/10.5121/csit.2020.100701

Leo, S. (2023, August 31). Effective Prompt Engineering Guide to Scale ChatGPT API for Sentiment Analysis on Multiple Survey Comments. *Medium*. https://shilpa-leo.medium.com/effective-prompt-engineering-guide-to-scale-chatgpt-api-for-sentiment-analysis-on-multiple-survey-c6d3dc924e74

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Meaden, J., Sturdivant, M., & Theys, E. R. (2024). Harnessing the Power of Generative AI through Effective Prompt Engineering [Master Tutorial]. *Society for Industrial and Organizational Psychology Annual Conference*, Chicago, IL, United States.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

OpenAI. (n.d.). Documentation overview. *OpenAI*. https://platform.openai.com/docs/overview

OpenAI. (n.d.). Fine-tuning guide. *OpenAI*. https://platform.openai.com/docs/guides/fine-tuning

OpenAI. (n.d.). GPT-4 fine-tuning overview. *OpenAI*. https://openai.com/index/gpt-4o-fine-tuning

OpenAI. (n.d.). Prompt engineering guide. *OpenAI*. https://platform.openai.com/docs/guides/prompt-engineering

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135. https://doi.org/10.1561/1500000011

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023). Towards coding social science datasets with language models. *arXiv preprint arXiv:2306.02177*. https://arxiv.org/abs/2306.02177

Stone, D. L., Deadrick, D. L., Lukaszewski, K. M., & Johnson, K. R. (2015). The influence of technology on the future of human resource management. *Human Resource Management Review*, 25(2), 216-231. https://doi.org/10.1016/j.hrmr.2015.01.002

Thomas, K. (Co-chair), Kruse, K., Carson, B., Callahan, K., Cabe, M (2024). How Innovative Technology Is Shifting Leadership Development [Panel presentation]. *Society for Industrial and Organizational Psychology Annual Conference*, Chicago, IL, United States.

Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing and Health Sciences*, 15(3), 398-405. https://doi.org/10.1111/nhs.12048

Appendix: Prompt Chains and Results Tables

—

Exploratory Analysis

Table 9. Exploratory Analysis Prompt Chain.

Prompt Number	Text
1	{upload data}
	The data in the file is from an employee survey and it consists of a unique ID for each respondent in column 1 followed by that respondent's written comment in column 2.
	Conduct an Overall sentiment analysis providing the number of positive, neutral, negative, mixed comments and place the results in a table; column 1 of the table should be "Sentiment" and from row 2 the sentiment categories positive, negative, neutral, and mixed; column 2 of the table should be the number of comments that fall into each sentiment category.
2	Create a table representing each employee's unique ID, their response, and sentiment category.

Table 10. Exploratory Analysis Confusion Matrix.

	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	10	0	37	0
Actual Negative	2	7	56	2
Actual Neutral	0	0	19	0
Actual Mixed	0	1	4	0

Table 11. Exploratory Analysis Classification Table.

Class	Precision	Recall	F1-Score	AI Count	Human Count*
Positive	0.83	0.21	0.34	12	47
Negative	0.88	0.10	0.19	8	67
Neutral	0.16	1.00	0.28	116	19
Mixed	0.00	0.00	0.00	2	5

Note: *The “Human Count” column refers to the number of comments per sentiment category as determined by the human SME.

—

Iteration 1

Table 12. Iteration 1 Prompt Chain.	
Prompt Number	Text
1	{upload data}
	Please conduct a sentiment analysis on the provided comments, paying close attention to context, negations, sarcasm, and any signs of mixed feelings. Use the following classification guidelines:
	Label a comment as 'positive' if it expresses clear satisfaction, joy, or praise throughout. Be cautious of comments that seem positive at first but contain subtle criticisms; such comments should not be classified as purely positive. Label a comment as 'negative' if it expresses frustration, disapproval, or dissatisfaction. Watch for sarcastic or ironic language, as these can flip the meaning of otherwise neutral-sounding phrases. For example, 'Oh sure, the management really cares' should be labeled as negative. Label a comment as 'neutral' if it contains primarily factual information without strong emotion. These comments report conditions or provide context without expressing any clear satisfaction or frustration. Label a comment as 'mixed' if it contains both positive and negative aspects. Pay attention to comments that praise one aspect but criticize another, such as when someone expresses love for their team but frustration with leadership. Mixed comments may also contain contradictions or nuanced sentiments.

Table 13. Iteration 1 Confusion Matrix.				
Actual	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	28	12	66	9
Actual Negative	11	31	51	12
Actual Neutral	0	3	8	1
Actual Mixed	6	24	11	19

Table 14. Iteration 1 Classification Table.

Class	Precision	Recall	F1-Score	AI Count	Human Count*
Positive	0.62	0.24	0.35	136	115
Negative	0.44	0.30	0.35	70	105
Neutral	0.06	0.67	0.11	45	12
Mixed	0.46	0.32	0.38	41	59

Note: *The “Human Count” column refers to the number of comments per sentiment category as determined by the human SME.

—

Iteration 2

Table 15. Iteration 2 Prompt Chain.	
Prompt Number	Text
1	Please conduct a sentiment analysis on the provided comments, paying close attention to context, negations, sarcasm, and any signs of mixed feelings. Use the following classification guidelines: Label a comment as 'positive' if it expresses clear satisfaction, joy, or praise throughout. Be cautious of comments that seem positive at first but contain subtle criticisms; such comments should not be classified as purely positive. Label a comment as 'negative' if it expresses frustration, disapproval, or dissatisfaction. Watch for sarcastic or ironic language, as these can flip the meaning of otherwise neutral-sounding phrases. For example, 'Oh sure, the management really cares' should be labeled as negative. Label a comment as 'neutral' if it contains primarily factual information without strong emotion. These comments report conditions or provide context without expressing any clear satisfaction or frustration. Label a comment as 'mixed' if it contains both positive and negative aspects. Pay attention to comments that praise one aspect but criticize another, such as when someone expresses love for their team but frustration with leadership. Mixed comments may also contain contradictions or nuanced sentiments. Additional Guidance: Handle negations carefully (e.g., 'I would not recommend' is a negative statement despite the neutral wording). Ensure that negated or contradictory statements are accurately reflected in the sentiment classification. For comments with sarcasm, look for clues that reverse the apparent sentiment. If a comment appears positive but is intended sarcastically, label it as negative or mixed based on the overall tone. Focus on precision when classifying comments. If there's any ambiguity between positive and mixed, default to labeling the comment as mixed. Apply the same classification standards across all comments to maintain consistency and closely follow the human-rater style, ensuring accurate and context-sensitive sentiment detection.
2	{upload data} Compare your sentiment analysis results to the "correct" answers from the human coders.

Table 16. Iteration 2 Confusion Matrix.

	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	77	1	25	12
Actual Negative	13	21	49	22
Actual Neutral	5	1	5	1
Actual Mixed	14	7	10	28

Table 17. Iteration 2 Classification Table.

Class	Precision	Recall	F1-Score	AI Count	Human Count*
Positive	0.71	0.67	0.69	109	115
Negative	0.70	0.20	0.31	30	105
Neutral	0.06	0.42	0.10	89	12
Mixed	0.44	0.47	0.46	63	59

Note: *The “Human Count” column refers to the number of comments per sentiment category as determined by the human SME.

—

Iteration 3

Table 18. Iteration 3 Prompt Chain.

Prompt Number	Text
1	{upload data} Please conduct a sentiment analysis on the provided comments, paying close attention to context, negations, sarcasm, and any signs of mixed feelings. Use the following classification guidelines: Classification Guidelines: Label a comment as 'positive' if more than 70% of the comment expresses clear satisfaction, joy, or praise throughout. Be cautious of comments that seem positive but contain subtle criticisms; such comments should not be classified as purely positive unless the positive sentiment clearly dominates. Label a comment as 'negative' if more than 70% of the comment expresses frustration, disapproval, or dissatisfaction. Pay close attention to sarcastic or ironic language, as it can reverse the meaning of seemingly neutral or positive phrases (e.g., 'Oh sure, the management really cares' should be labeled as negative). Label a comment as 'neutral' if it contains primarily factual information (over 70% of the comment) without strong emotions. Neutral comments typically report conditions or provide context without expressing clear satisfaction or frustration. Label a comment as 'mixed' if it contains both positive and negative aspects, with neither tone clearly dominating or if the balance between positive and negative content falls between 30% and 70%. Pay attention to nuanced sentiments where praise for one aspect is countered by criticism of another. In cases of ambiguity, default to labeling the comment as mixed. Additional Guidance: Proportion rule: Classify the comment based on the overall proportion of positive, negative, and neutral tones. If one sentiment (positive or negative) clearly dominates (i.e., above 70%), assign that sentiment. If sentiments are relatively balanced, label the comment as mixed. Negations: Handle negations carefully. For example, phrases like 'I would not recommend' are negative despite neutral wording. Ensure that contradictions and negated statements are appropriately accounted for in the sentiment classification. Sarcasm and irony: Look for clues that reverse the apparent sentiment. Comments that appear positive but are sarcastic should be labeled as negative or mixed depending on the tone. Precision and consistency: Maintain consistency by applying the same classification standards across all comments. Focus on accurately reflecting the intended sentiment, and always weigh the proportion of positive vs. negative content when determining the final classification.
2	{upload data} Compare your sentiment analysis results to the "correct" answers from the human coders.

Table 19. Iteration 3 Confusion Matrix.

	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	40	26	46	3
Actual Negative	1	56	39	9
Actual Neutral	1	5	6	0
Actual Mixed	2	35	10	13

Table 20. Iteration 3 Classification Table.

Class	Precision	Recall	F1-Score	AI Count	Human Count*
Positive	0.91	0.35	0.50	122	115
Negative	0.46	0.53	0.49	101	105
Neutral	0.06	0.50	0.11	44	12
Mixed	0.52	0.22	0.31	25	59

Note: *The “Human Count” column refers to the number of comments per sentiment category as determined by the human SME.

—

Iteration 4

Table 21. Iteration 4 Prompt Chain.

Prompt Number	Text
1	{upload data} Here is a dataset that includes the columns 'UserID,' 'Comment,' and 'Sentiment Category.' The sentiment category of each comment was manually coded by human raters as either 'positive,' 'negative,' 'neutral,' or 'mixed.' Please learn from the examples in this dataset how the comments are classified according to these sentiment categories. When analyzing new comments, use the patterns from the dataset to inform your classification. Ensure that you consider context, negations, sarcasm, and any signs of mixed sentiment in your analysis, and apply a similar classification style to maintain consistency with the human-rater approach.
2	{upload data} Use the dataset in the previous prompt to classify these new comments following the human-rater style.

Table 22. Iteration 4 Confusion Matrix.

	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	28	7	0	0
Actual Negative	1	32	0	0
Actual Neutral	0	1	0	3
Actual Mixed	4	13	0	0

Table 23. Iteration 4 Classification Table.					
Class	Precision	Recall	F1-Score	AI Count	Human Count*
Positive	0.85	0.80	0.82	33	35
Negative	0.60	0.97	0.74	53	33
Neutral	0.00	0.00	0.00	0	4
Mixed	0.00	0.00	0.00	3	17

Note: *The “Human Count” column refers to the number of comments per sentiment category as determined by the human SME.

—

Iteration 5

Table 24. Iteration 5 Prompt Chain.	
Prompt Number	Text
1	{upload data} Here is a dataset that includes the columns 'UserID,' 'Comment,' and 'Sentiment Category.' The sentiment category of each comment was manually coded by human raters as either 'positive,' 'negative,' 'neutral,' or 'mixed.' Please learn from the examples in this dataset how the comments are classified according to these sentiment categories. When analyzing new comments, use the patterns from the dataset to inform your classification. Ensure that you consider context, negations, sarcasm, and any signs of mixed sentiment in your analysis, and apply a similar classification style to maintain consistency with the human-rater approach. For context, the employees were responding to the open-ended question: “If a close friend of yours were looking for a job, would you recommend your [location] as a good place to work? Why or why not?”
2	{upload data} Use the dataset in the previous prompt to classify these new comments following the human-rater style.

Table 25. Iteration 5 Confusion Matrix.

	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	34	1	0	0
Actual Negative	8	25	0	0
Actual Neutral	3	1	0	0
Actual Mixed	10	7	0	0

Table 26. Iteration 5 Classification Table.

Class	Precision	Recall	F1-Score	AI Count	Human Count*
Positive	0.62	0.97	0.76	55	35
Negative	0.74	0.76	0.75	34	33
Neutral	0.00	0.00	0.00	0	4
Mixed	0.00	0.00	0.00	0	17

Note: *The “Human Count” column refers to the number of comments per sentiment category as determined by the human SME.

—

Iteration 6

Table 27. Iteration 6 Prompt Chain.

Prompt Number	Text
1	Please conduct a sentiment analysis on the provided comments, paying close attention to context, negations, sarcasm, and any signs of mixed feelings. Use the following classification guidelines: Label a comment as 'positive' if: More than 70% of the comment expresses clear satisfaction, joy, or praise. Pay special attention to phrases that convey support, encouragement, motivation, recognition, or team cohesion such as "feel supported," "encouraged," "motivated," or "valued," and assign greater weight to these expressions. Additionally, detect emphatic words like "definitely," "absolutely," and "always" used in a positive context, as they indicate a stronger positive sentiment. In cases where these phrases and words are present without substantial negative elements, classify the comment as positive. For context, the employees were responding to the open-ended question: “If a close friend of yours were looking for a job, would you recommend your [location] as a good place to work? Why or why not?” If they say “yes” or “yes they would recommend” the workplace or location, that would be considered a positive statement. If these positive elements are present without substantial negative sentiment, classify the comment as positive. Capture Subtle Positive Indicators: Assign weight to phrases that express general approval, such as "good place," "satisfactory," "decent environment," or "okay experience" even if they lack emphatic language. These indicators still contribute to a positive sentiment. Ensure that positive classifications account for any negative aspects or uncertainties: If a comment includes a positive sentiment alongside conditional phrases like "if things improve" or "depends on," consider the uncertainty and weigh it appropriately before finalizing the classification. If the comment ultimately reflects a positive experience, the overall sentiment should be positive, even if the recommendation is conditional.

Label a comment as 'negative' if:
More than 70% of the comment expresses frustration, disapproval, or dissatisfaction.
Pay close attention to sarcastic or ironic language, as it can reverse the meaning of seemingly neutral or positive phrases (e.g., "Oh sure, the management really cares" should be labeled as negative).
For context, the employees were responding to the open-ended question: “If a close friend of yours were looking for a job, would you recommend your [location] as a good place to work? Why or why not?” If they say “no” they would not recommend the workplace or location, that would be considered a negative statement.
Multiple mentions of negative aspects, criticisms, or descriptions of challenges should also be considered negative statements.
Pay special attention to strong negations, such as "absolutely not," "definitely not," "never," or "would not recommend." These phrases are strong indicators of negative sentiment and should be given significant weight in the analysis. If a comment includes such a phrase, classify it as negative unless there is an overwhelming and explicit positive aspect that dominates the comment.
Ensure that negation words like "not," "don't," "would not," and "never" are used to correctly reverse the intended sentiment of otherwise positive phrases, and consider repetition or emphasis of negative points as a critical factor in determining overall sentiment.
Assign significant weight to key negative phrases like "overworked," "underpaid," "stressful," "toxic," and similar descriptors that indicate poor conditions.

Accumulate negative phrases: If a comment includes multiple negative descriptors about workload, compensation, or stress, classify it as negative, as these cumulatively indicate dissatisfaction. Phrases like "employees are overworked, underpaid, and stressed" should strongly push the sentiment to negative.
If there are repeated criticisms or descriptions of negative conditions, these should take precedence even if there are other neutral statements.
Consider negative aspects even when positive phrases are present: If a negative recommendation is followed by a positive statement (e.g., "I would not recommend working here, but I think things are improving"), prioritize the negative sentiment to reflect the current dissatisfaction.
If the comment ultimately reflects a negative experience, the overall sentiment should be negative, even if the recommendation is conditional.
Label a comment as 'neutral' if:
It contains primarily factual information (over 70%) without strong emotions.
Neutral comments typically report conditions or provide context without expressing satisfaction or frustration. Be cautious of comments that seem neutral but contain criticisms; such comments should not be classified as purely neutral unless the neutral sentiment clearly dominates. Examples include statements like "it depends on the department" without further evaluation.
In cases of ambiguity, default to labeling the comment as neutral.
Conditional statements like "it depends on" or "not sure yet" may indicate uncertainty, which should influence the classification towards neutral if no strong emotions are present.

Label a comment as 'mixed' if:
The comment contains both positive and negative aspects, with neither clearly dominating.
Pay attention to nuanced sentiments where praise for one aspect is countered by criticism of another.

Be careful with comments containing a negative recommendation followed by a conditional or future-oriented positive aspect. In such cases, prioritize the current negative sentiment and classify the comment as negative.
Ensure balanced evaluation: If a comment contains both strong positive and negative elements, carefully weigh their proportion. If neither sentiment clearly outweighs the other, classify it as mixed.
Look for Equal Emphasis: Ensure that a single negative statement does not overpower an otherwise balanced comment. If both positive and negative sentiments are presented without one clearly dominating, classify the comment as mixed.

Additional Guidance:
Proportion Rule: Classify the comment based on the overall proportion of positive, negative, and neutral tones. If one sentiment (positive or negative) clearly dominates (i.e., above 70%), assign that sentiment. If sentiments are relatively balanced, label the comment as mixed.
Negations: Handle negations carefully. Recognize negation words like "don't," "not," "would not," etc., that reverse any otherwise positive sentiment. Phrases like "I would not recommend" are negative despite neutral wording. Ensure that contradictions and negated statements are appropriately accounted for in the sentiment classification.
Sarcasm and Irony: Look for clues that reverse the apparent sentiment. Comments that appear positive but are sarcastic should be labeled as negative or mixed depending on the tone.
Precision and Consistency: Maintain consistency by applying the same classification standards across all comments. Focus on accurately reflecting the current intended and dominant sentiment, and always weigh the proportion of positive vs. negative content when determining the final classification.
Enhanced Handling of Strong Phrases:
Assign higher weight to strong negative phrases such as "absolutely not" and "never," and apply boosts to repeated negative indicators.
Ensure that multiple negative descriptors (e.g., "overworked," "stressful") collectively push the sentiment towards negative, even if some neutral elements are present.
Avoid Over-Classifying Mixed Comments as Negative: If a comment contains negative statements that are mitigated by positive aspects (e.g., "despite challenges, there are improvements"), ensure that these comments are properly classified as mixed rather than defaulting to negative unless the negative tone is overwhelming.

Table 28. Iteration 6 Confusion Matrix.

	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	58	1	55	1
Actual Negative	22	33	34	16
Actual Neutral	0	0	12	0
Actual Mixed	28	4	16	12

Table 29. Iteration 6 Classification Table.

Class	Precision	Recall	F1-Score	AI Count	Human Count*
Positive	0.51	0.69	0.59	108	115
Negative	0.70	0.64	0.67	38	105
Neutral	0.00	0.00	0.00	117	12
Mixed	0.00	0.00	0.00	29	59

Note: *The “Human Count” column refers to the number of comments per sentiment category as determined by the human SME.

—

Iteration 7

Table 30. Iteration 7 Prompt Chain.	
Prompt Number	Text
1	{upload data} Here is a dataset that includes the columns 'UserID,' 'Comment,' and 'Sentiment Category.' The sentiment category of each comment was manually coded by human raters as either 'positive,' 'negative,' 'neutral,' or 'mixed.' Please learn from the examples in this dataset how the comments are classified according to these sentiment categories. When analyzing new comments, use the patterns from the dataset to inform your classification. Ensure that you consider context, negations, sarcasm, and any signs of mixed sentiment in your analysis, and apply a similar classification style to maintain consistency with the human-rater approach.
2	Use the dataset in the previous prompt to classify these new comments following the human-rater style. Use the following classification guidelines: Label a comment as 'positive' if: More than 70% of the comment expresses clear satisfaction, joy, or praise. Pay special attention to phrases that convey support, encouragement, motivation, recognition, or team cohesion such as "feel supported," "encouraged," "motivated," or "valued," and assign greater weight to these expressions. Additionally, detect emphatic words like "definitely," "absolutely," and "always" used in a positive context, as they indicate a stronger positive sentiment. In cases where these phrases and words are present without substantial negative elements, classify the comment as positive. For context, the employees were responding to the open-ended question: “If a close friend of yours were looking for a job, would you recommend your [location] as a good place to work? Why or why not?” If they say “yes” or “yes they would recommend” the workplace or location, that would be considered a positive statement. If these positive elements are present without substantial negative sentiment, classify the comment as positive. Capture Subtle Positive Indicators: Assign weight to phrases that express general approval, such as "good place," "satisfactory," "decent environment," or "okay experience" even if they lack emphatic language. These indicators still contribute to a positive sentiment. Ensure that positive classifications account for any negative aspects or uncertainties: If a comment includes a positive sentiment alongside conditional phrases like "if things improve" or "depends on," consider the uncertainty and weigh it appropriately before finalizing the classification. If the comment ultimately reflects a positive experience, the overall sentiment should be positive, even if the recommendation is conditional.

Label a comment as 'negative' if:
More than 70% of the comment expresses frustration, disapproval, or dissatisfaction.
Pay close attention to sarcastic or ironic language, as it can reverse the meaning of seemingly neutral or positive phrases (e.g., "Oh sure, the management really cares" should be labeled as negative).
For context, the employees were responding to the open-ended question: “If a close friend of yours were looking for a job, would you recommend your [location] as a good place to work? Why or why not?” If they say “no” they would not recommend the workplace or location, that would be considered a negative statement.
Multiple mentions of negative aspects, criticisms, or descriptions of challenges should also be considered negative statements.
Pay special attention to strong negations, such as "absolutely not," "definitely not," "never," or "would not recommend." These phrases are strong indicators of negative sentiment and should be given significant weight in the analysis. If a comment includes such a phrase, classify it as negative unless there is an overwhelming and explicit positive aspect that dominates the comment.
Ensure that negation words like "not," "don't," "would not," and "never" are used to correctly reverse the intended sentiment of otherwise positive phrases, and consider repetition or emphasis of negative points as a critical factor in determining overall sentiment.
Assign significant weight to key negative phrases like "overworked," "underpaid," "stressful," "toxic," and similar descriptors that indicate poor conditions.
Accumulate negative phrases: If a comment includes multiple negative descriptors about workload, compensation, or stress, classify it as negative, as these cumulatively indicate dissatisfaction. Phrases like "employees are overworked, underpaid, and stressed" should strongly push the sentiment to negative.
If there are repeated criticisms or descriptions of negative conditions, these should take precedence even if there are other neutral statements.
Consider negative aspects even when positive phrases are present: If a negative recommendation is followed by a positive statement (e.g., "I would not recommend working here, but I think things are improving"), prioritize the negative sentiment to reflect the current dissatisfaction.
If the comment ultimately reflects a negative experience, the overall sentiment should be negative, even if the recommendation is conditional.
Label a comment as 'neutral' if:
It contains primarily factual information (over 70%) without strong emotions.
Neutral comments typically report conditions or provide context without expressing satisfaction or frustration. Be cautious of comments that seem neutral but contain criticisms; such comments should not be classified as purely neutral unless the neutral sentiment clearly dominates. Examples include statements like "it depends on the department" without further evaluation.
In cases of ambiguity, default to labeling the comment as neutral.
Conditional statements like "it depends on" or "not sure yet" may indicate uncertainty, which should influence the classification towards neutral if no strong emotions are present.
Label a comment as 'mixed' if:
The comment contains both positive and negative aspects, with neither clearly dominating.
Pay attention to nuanced sentiments where praise for one aspect is countered by criticism of another.
Be careful with comments containing a negative recommendation followed by a conditional or future-oriented positive aspect. In such cases, prioritize the current negative sentiment and classify the comment as negative.

Ensure balanced evaluation: If a comment contains both strong positive and negative elements, carefully weigh their proportion. If neither sentiment clearly outweighs the other, classify it as mixed.
Look for Equal Emphasis: Ensure that a single negative statement does not overpower an otherwise balanced comment. If both positive and negative sentiments are presented without one clearly dominating, classify the comment as mixed.
Additional Guidance:
Proportion Rule: Classify the comment based on the overall proportion of positive, negative, and neutral tones. If one sentiment (positive or negative) clearly dominates (i.e., above 70%), assign that sentiment. If sentiments are relatively balanced, label the comment as mixed.
Negations: Handle negations carefully. Recognize negation words like "don't," "not," "would not," etc., that reverse any otherwise positive sentiment. Phrases like "I would not recommend" are negative despite neutral wording. Ensure that contradictions and negated statements are appropriately accounted for in the sentiment classification.
Sarcasm and Irony: Look for clues that reverse the apparent sentiment. Comments that appear positive but are sarcastic should be labeled as negative or mixed depending on the tone.
Precision and Consistency: Maintain consistency by applying the same classification standards across all comments. Focus on accurately reflecting the current intended and dominant sentiment, and always weigh the proportion of positive vs. negative content when determining the final classification.
Enhanced Handling of Strong Phrases:
Assign higher weight to strong negative phrases such as "absolutely not" and "never," and apply boosts to repeated negative indicators. Ensure that multiple negative descriptors (e.g., "overworked," "stressful") collectively push the sentiment towards negative, even if some neutral elements are present.
Avoid Over-Classifying Mixed Comments as Negative: If a comment contains negative statements that are mitigated by positive aspects (e.g., "despite challenges, there are improvements"), ensure that these comments are properly classified as mixed rather than defaulting to negative unless the negative tone is overwhelming.

Table 31. Iteration 7 Confusion Matrix.

	Predicted Positive	Predicted Negative	Predicted Neutral	Predicted Mixed
Actual Positive	35	0	0	0
Actual Negative	10	23	0	0
Actual Neutral	1	3	0	0
Actual Mixed	14	3	0	0

Table 32. Iteration 7 Classification Table.

Class	Precision	Recall	F1-Score	AI Count	Human Count*
Positive	0.51	0.69	0.59	60	35
Negative	0.70	0.64	0.67	29	33
Neutral	0.00	0.00	0.00	0	4
Mixed	0.00	0.00	0.00	0	17

Note: *The “Human Count” column refers to the number of comments per sentiment category as determined by the human SME.

Learn more about our capabilities

HRTec has experienced personnel standing by to walk your team through the entire organizational assessment process or any area of your choice

- Assessment Consultation
- Surveys / Assessments
- Secure Hosting
- Data Analytics
- Post Analysis Reporting
- Comment Analysis
- Focus Groups / Interviews

How to cite this report:

Pagan, A.D. (Presenter), Pagan, A.M., & Steinhauser, E.F. (2025, April 3). Harnessing Generative AI for Enhanced Sentiment Analysis in Organizational Settings [Poster presentation]. Society for Industrial and Organizational Psychology Annual Conference, Denver, CO, United States. https://www.surveyqwik.com

HRTec
Assessment System (HAS)

Purpose built compliance and technical solutions that assist organizations with achieving their unique business and mission goals and objectives.

info@surveyqwik.com

866-933-4999

surveyqwik.com