RESEARCH INSIGHTS

AI-Assisted Text Analysis for Coaching Evaluation

Exploring the Use of a Common AI Tool for Theming Text-Based Datasets

By Katelyn McCoy, Tanner Landolt, Ezekiel Welsh, and Caroline DeStefano

Center for Creative Leadership

https://doi.org/10.35613/ccl.2025.2064

CONTENTS

Executive Summary

Over the past few years, publicly accessible AI tools have taken the world by storm. There seems to be a tool for everything these days, from helping with grocery lists to conducting data analyses. At CCL, we've been exploring how AI can help us better serve our clients, including how we can bring data-driven insights to more clients, faster. Particularly in the case of our evaluation data, the process of synthesizing text-based results in a meaningful way can be time-consuming, yet it's valuable for adding nuance to quantitative results. AI could be the answer to a more efficient meaning-making process for large amounts of text-based data, but how can we be sure its output is accurate and useful?

The following study came from a need to test just that. When leaders take the time to write about their experiences in our feedback surveys, we want to be sure there's integrity in our process of summarizing what they said. With over 6,000 responses available from one of our open-ended evaluation survey questions, we decided to test how one publicly available large language model could assist with theming the data.

To do that, we engaged both humans and OpenAI's GPT4 model in the same task – thematically coding leaders' responses. Then, we compared the results. While our human coders and GPT4 varied greatly in their efficiency (of course GPT4 was much faster), we had three key takeaways.

First, GPT4's output matched the consensus of our human coders 55-65% of the time when all parties were asked to select one theme from a GPT4-generated set to summarize each response. This is only slightly lower than the rate at which our human coders matched each other pre-consensus (60-70%) and suggests GPT4 can be nearly as useful as a human when initially coding large numbers of responses.

Second, when we gave both parties the opportunity to select more than one theme from the set to summarize responses, at least one of GPT4's themes matched that of the human consensus 85% of the time, but GPT4 selected more than twice as many themes as our human coders. In this case, our human coders demonstrated more discretion and nuance by selecting only those themes that most closely matched the meaning behind the response, and GPT4 output a lot more noise.

Finally, we engaged our expert coders to refine the set of themes GPT4 initially provided before including them as input to GPT4's multi-theme coding task. By doing this, we reduced the number of themes GPT4 output by nearly half, while maintaining its accuracy at around 75%. From this, we learned that engaging in strategic human intervention (i.e., a human-in-the-loop process) enabled us to increase the utility of AI for this data analysis task, striking a balance between efficiency and accuracy.

Introduction

At the Center for Creative Leadership, we're interested in understanding how activities we include in our solutions impact leaders' development. One way we investigate this is by engaging in ongoing, robust feedback processes, which range from immediate evaluation to long-term impact measurement. Often, we find that leaders' written-in or spoken feedback provides a much richer picture of how a particular activity impacted their development than the ratings they provide on a feedback survey, no matter how creative we are with the questions we ask (Bukin & Schneider, 2022). This is particularly true when it comes to one-on-one coaching included in our leadership development programs (called *integrated coaching*), which is highly personalized.

Both in our core leadership training programs (open enrollment for individual leaders from different organizations) and programs customized to the unique needs of particular organizations, we often include integrated coaching for two reasons.

1. To help leaders reflect on, internalize, and apply what they've learned from other developmental activities (including assessments)
2. To support leaders through identifying and pursuing goals, driven by their personal leadership challenges

The goal of integrated coaching is to help participants progress from their current leadership practices to more effective future practices, and that can look different for each individual leader who experiences it. Because of that, we tend to receive varied feedback about how coaching benefited leaders, and that feedback is included in their written-in responses rather than their satisfaction ratings. As rich as this written feedback can be, analyzing large amounts of qualitative (unstructured, text-based) data is time-consuming. To truly help our clients understand the impact of this activity on leaders, it's important that we're able to accurately summarize qualitative responses in a way that helps decision-makers cut right to the insights they contain, and quickly.

In their blog post, Bukin and Schneider (2022) describe a seven-step process CCL often utilizes to leverage natural language processing technology in synthesizing large amounts of participant responses. Other methods include entirely human-led or technology-assisted thematic coding (Williams & Moser, 2019), as well as newer, experimental methods that rely entirely on automated output (Khan et al., 2024). Without specially trained automated tools, the tradeoff is often one of resources (time and human expertise) versus accuracy, but with advances in publicly available AI happening every day, we wondered how we could find a balance between the two.

With over 6,000 open-ended feedback survey responses from the last few years about how integrated coaching impacted leaders' development, we experimented with ways to engage AI in synthesizing key themes. In this paper, we explore the steps we took to better understand how AI could identify themes from real feedback data and assign those themes to individual comments.

Using AI to Theme Responses to Feedback Surveys

We opted to use Azure OpenAI's API (accessed via Python) because it would allow us to leverage a closed (or private) version of OpenAI's LLM (large language model). Although we removed all identifying data from the dataset before engaging in the methods described below, we took this extra precaution to ensure the security of our customer data, while also closely mimicking the output that might be produced by ChatGPT, a more easily accessible interface for many practitioners and evaluators. Throughout this paper, we refer to the AI model used as "GPT4" because we engaged with the GPT-4 model, which at the time of our study offered a practical mix of sophistication and accessibility.

It's worth noting that most publicly available AI models have documentation and guidance for use on their websites. In the case of this study, you can visit the OpenAI platform information site to learn more about the strengths and limitations of the model we used. We acknowledge that in the time it has taken to share our study results, several other models have been released across the AI innovation landscape, including open-source ones, that might fare better or worse at this task. For the purposes of this study, we did not compare models. Rather, our goal was to compare one popular model's ability to efficiently and accurately theme participant comments with that of human coders.

Identifying Our Samples

CCL provides integrated coaching in several different contexts globally, so we narrowed it down to the following three samples for this study:

1. **Open Enrollment Sample:** This sample represents one-on-one coaching that takes place during our individual leadership training programs (what we call "open enrollment").
2. **Custom Sample:** This sample represents one-on-one coaching that takes place during a program tailored to an organization's unique context, challenges, and culture (what we call "custom").
3. **Extended Custom Sample:** This sample represents one-on-one coaching included in a custom program that takes place at a different time from the other program activities. Although we could have included these responses in the Custom Sample, we separated the two because we suspected that the different timing of the coaching would yield different themes.

In each of these contexts, our feedback surveys included the same open-ended prompt: *Please comment on how coaching impacted your leadership development experience.* We took the following basic steps to select the text data included in each sample.

First, because we were most interested in the aspects of coaching that positively impacted leaders' development, we included only responses from participants who indicated they were satisfied with their coaching experience. Our cutoffs for this were a four or above on a five-point satisfaction scale, or an eight or above on a zero-to-ten likelihood-to-recommend scale. After this, we systematically removed comments with fewer than four words so that the remaining ones were more likely to yield high-quality insights into the impact of coaching. From the resulting subset, we then randomly selected 500 comments from each context to create our three representative samples.

Descriptive statistics for the comments included in our samples are provided in this paper's Appendix (See Table 1).

Engaging GPT4 in Thematic Coding

We then engaged GPT4 to identify the themes that were present within each sample. To do this, we appended the 500 comments identified for a particular sample to GPT4 in the form of a corpus, or large block of text. This meant that, initially, we did not ask GPT4 to distinguish between individual comments, but rather to consider all the comments in the sample as a unit. We then used the following prompt to request that the model return

a list of common themes based on the text provided for that sample.

> ***I'm looking at responses to a survey prompt that reads: "Please comment on how coaching impacted your leadership development experience." Each message response below is a comment from a leader who experienced coaching as part of a leadership development program. Provide a list of common themes based on these comments:***

This yielded 45 total themes, 15 emerging from each sample. GPT4's initially generated themes and their definitions for each sample are included in the Appendix (See Table 2).

At this point, our research team reviewed all 45 themes to ensure they appeared accurate given the data we had provided to GPT4. As expected, we noticed several of the theme names and definitions were similar across the three samples, which provided preliminary evidence that GPT4 had accurately summarized the data. Additionally, several themes that stood out as unique to a particular sample seemed to align with the sample's context. For example, personalization stood out as a theme in our Open Enrollment Sample and no others, which reflects the reality that the integrated coaching might provide a better opportunity for personal attention in programs that are more standardized, as our core training programs are. Satisfied with the themes, we moved on to our next step to determine just how representative each theme was of the comments contained in each sample.

Our approach is summarized in Figure 1 below.

THEMATIC CODING PROCESS ENGAGING GPT4 AND HUMAN CODERS

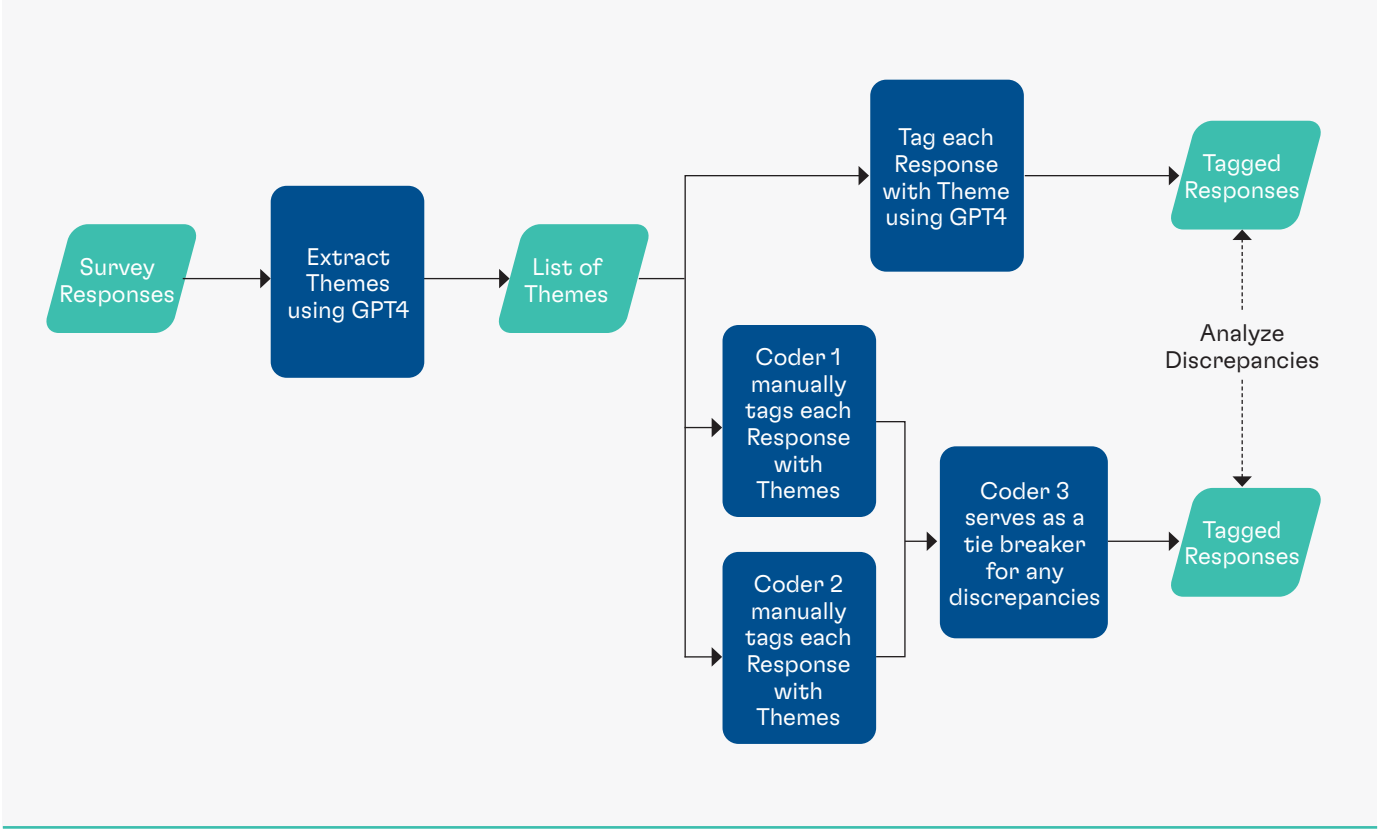


FIGURE 1

Single Theme Tagging

Our next step was to see if GPT4 could accurately assign (or tag) the 15 themes it had extracted to the 500 individual comments in each sample. Our goal was to understand the prevalence of the themes within each sample and compare the GPT4-based theme tagging to that of subject-matter-expert coders.

For each sample, we instructed GPT4 to tag every individual comment with one of its provided themes, or if no themes were a match, to return the phrase "I don't know." To do this, we provided GPT4 with the following prompt, followed by two inputs: 1) its own previously generated themes and definitions and 2) an individual comment. Using Python, we then looped through this process for each of the 500 comments within each sample.

> ***I'm looking at a response to this prompt: "Please comment on how coaching impacted your leadership development experience."***
>
> ***Below is a comment from a leader who experienced coaching as part of a leadership development program, as well as a list of themes.***
>
> ***Return the one theme from the list that best categorizes the comment. Do not return the theme definition. Only return the theme like it is listed with no added punctuation. Do not hallucinate a different theme. If you do not know, return "I don't know".***

Initially, we observed that GPT4 would hallucinate the same theme to multiple responses consecutively, even if the theme did not reflect the comment as well as other themes did. It was obvious that GPT4 was confounding previously tagged comments with each additional one as it progressed through the list and was likely due to user error in our interaction with the API. To avoid this, we adjusted our code to initiate a new session with GPT4 for each consecutive comment in the list, repeating this process for all 500 comments in a loop. Using this as our final method, we were able to generate more accurate single-theme tags for each comment in each of our three samples. This process was fully automated and took roughly 30 minutes to complete for each sample.

Human Review

With our GPT4-tagged samples in hand, we then turned to our team of four (human) subject-matter-expert coders, all trained professionals in the field of leadership development. Without showing them the tags GPT4 had produced, we provided them with the GPT4-generated themes and asked them to complete their own thematic coding for the Open Enrollment Sample and the Custom Sample. We saved the Extended Custom Sample for a next step in the study – testing GPT4's ability to tag multiple themes instead of just one.

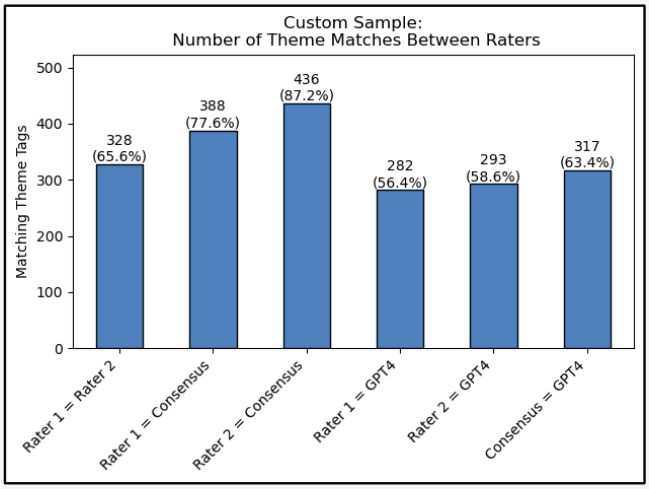
For each of these two samples, we engaged two initial coders and a third consensus coder, or tiebreaker. We instructed our initial coders to independently read the list of themes, then tag the single most relevant theme for each of the 500 comments. If there was no clear theme present within the comment, we instructed coders to tag "I don't know". If there was a more relevant theme present in a comment that was not within the themes provided by GPT4, coders were instructed to tag "Theme – Other" and provide a brief description of the theme. Finally, our consensus coder was instructed to finalize the themes tagged to each comment, settling any disputes between the previous coders by choosing the theme they thought was most representative of each comment.

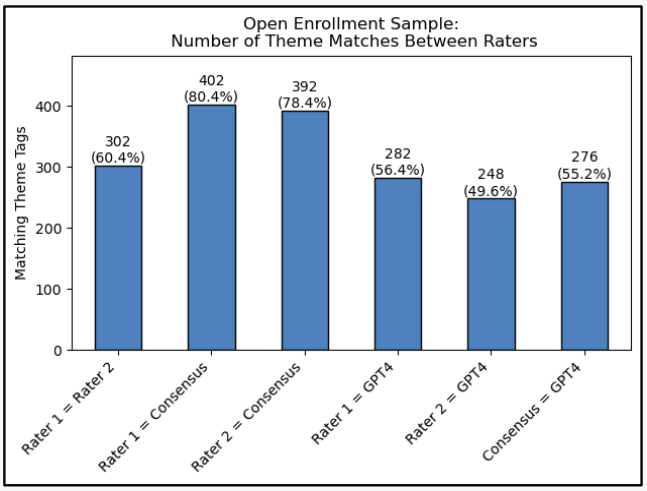
Results

We compared the single themes tagged by our two human coders, our consensus coder, and GPT4 to determine how GPT4 behaved as a coder compared to humans. First, we compared the frequency with which each combination of coders returned the same theme for a comment. Overall, decisions between independent coders matched between 60.4% (Coder 1 & Coder 2; Open Enrollment Sample) and 65.6% (Coder 1 & Coder 2; Custom Sample) of the time. Across both samples, the tags of our independent human coders matched each other slightly more frequently than either coder or the consensus coder matched the GPT4 output.

We calculated Cohen's Kappa (κ) for each combination of coders to better compare their levels of agreement (McHugh, 2012). This statistic measures agreement between two coders while accounting for possible random error introduced by the coding activity. It's calculated on a scale of 0-1, where above 0.5 is interpreted as higher than chance agreement and a

SINGLE THEME TAGGING RESULTS SAMPLES





higher number indicates more true agreement between coders. Throughout this study, our values generally fell between 0.5 and 0.7, all above chance, so we used this metric primarily to compare GPT4's performance to that of our human coders.

In our Custom Sample, the levels of agreement between two human coders and between the human consensus and GPT4 were the same ($\kappa = 0.586$), suggesting that GPT4 performed very similarly to the human coders. However, in our Open Enrollment Sample, human coders agreed slightly more frequently with each other ($\kappa = 0.558$) than GPT4 did with the human consensus ($\kappa = 0.505$).

A further examination of the data suggested that our human coders tended to agree slightly more because of their ability to discern the most appropriate theme and consider the context of the comment. The human coders generally tagged a singular appropriate theme more accurately when more than one theme *could be* present. As well, the human coders tended to more accurately assign themes in cases where certain key words misguided GPT4. In these cases, GPT4 tagged themes based on key words that had a different meaning when considered in the context of the comment, such that human coders outperformed GPT4. (See Appendix, Table 3).

It's worth noting that GPT4 also tended to assign themes when a human coder might have distinguished that the comment did not fit at all in the list of themes provided. For example, in the Open Enrollment sample the human consensus tagged 52 comments with "I don't know" and 6 comments with "Theme – Other", while GPT4 tagged only 15 comments total with "I don't know". (See Appendix, Table 4).

Multiple Theme Tagging

Following a debrief of the single-theme tagging task, we concluded that one reason why the high level of disagreement in the previous two samples (between human coders *and with* GPT4) was because more than one theme may have been equally appropriate to tag for a single comment. Although our human coders tended to choose the *most* appropriate theme, they pointed out that in many cases two or more themes may have still been accurate. In other words, the themes GPT4 had originally generated tended to overlap in meaning when applied to the comments. Because of this, we used our Extended Custom Sample to test the accuracy of GPT4 when we allowed it to tag *multiple themes* to a single comment.

We created a third GPT4 prompt to accomplish this task, which otherwise worked the same way as the single theme tagging. We provided the exact prompt we used in the appendix.

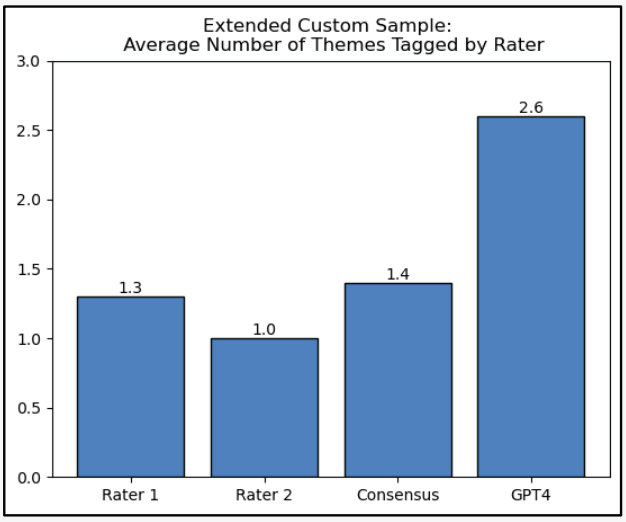
Human Review

As before, we asked two initial coders to independently assign themes to all 500 comments, but this time we told them to tag multiple themes to each comment where appropriate. Then, our consensus reviewer made the final decision about which themes to retain. Like the previous review task, we instructed coders to tag comments that where thematically related to something outside the provided list of themes with "Theme – Other" or, in cases were the comment was truly not related to any theme, "I don't know". We didn't place a limit on the number of themes tagged as long as they represented the comment.

Results

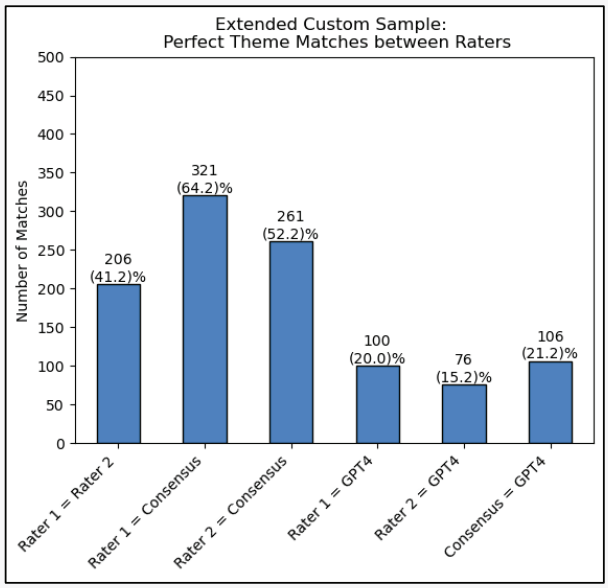
On average, GPT4 returned 2.6 themes for each comment in the sample, excluding comments that the model tagged "I don't know". By comparison, our human coders tagged almost half as many themes to each comment, with the consensus coder tagging an average of 1.4 themes. This reinforced our earlier observation that GPT4 was more likely to assign themes to comments than it was to exclude them, even in cases where our human reviewers determined the theme was inaccurate or simply not as relevant as others.

AVERAGE NUMBER OF THEMES TAGGED BY RATER



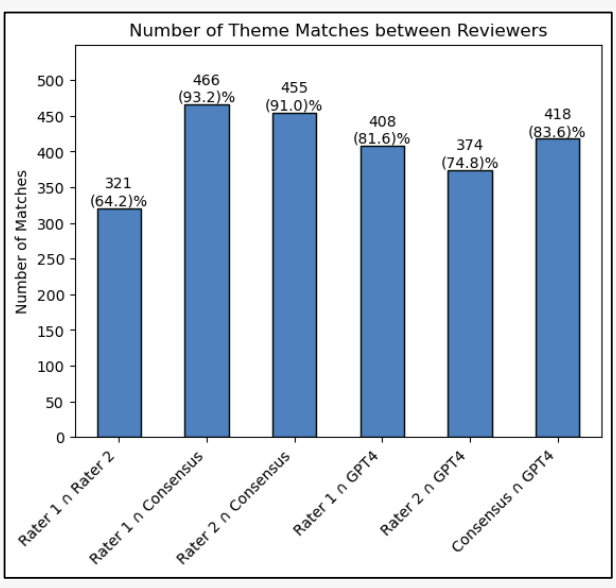
To assess agreement between the coders and GPT4, we first considered situations where the list of themes (as a whole) was an exact match. In this case, our human coders had lower agreement with each other than in the previous task. Coder 1 and Coder 2's provided list of themes was an exact match 41.2% of the time (κ = 0.357). The lists generated by consensus and GPT4 were exact matches much less frequently, only 21.6% of the time (κ = 0.126). This was primarily due to the number of themes GPT tagged, which made it less likely that its list would perfectly match the significantly shorter lists of themes represented by the human consensus.

PERFECT THEME MATCHES BETWEEN RATERS



Next, we examined how often *any* of the themes in the lists generated by the human coders and GPT4 matched. In this case, the lists generated by human coders contained a match 64.2% of the time, but the consensus and GPT4 lists contained a match 83.6% of the time. This suggests that when given the opportunity to tag multiple themes, GPT4 tended to tag enough of them that at least one matched the more discerning human consensus, while the others were not always accurate or nuanced representations of the meaning of the comment.

NUMBER OF THEME MATCHES BETWEEN REVIEWERS



Multiple Theme Tagging with Human Assistance

Our exercise of tagging multiple themes prompted us to take a closer look at the themes and definitions originally identified by GPT4. Our Extended Custom Sample, in particular, contained the most themes that were conceptually overlapping. For example, our human consensus tagged 39 comments with both "Self-Understanding" and "Increased Self-Awareness", and GPT4 tagged 101 comments with the same two themes. Because of this, we thought it would be valuable to engage in human intervention to clearly distinguish the themes we were providing as an input, with the end goal of improving GPT4's accuracy.

We engaged each of the researchers associated with this study in the task of refining the themes GPT4 had already identified. In this activity, we compared overarching thematic similarities across all three samples and distinguished those themes that appeared to be unique based on the context of each individual sample. As well, now that we'd established some evidence for the prevalence of each theme within and across each sample, we used that information to help guide our decision-making. For example, if a theme had been tagged to a higher number of comments, we made room for it in our refined list of themes. If it was only tagged to a few comments across all samples, then we did not include it in our refined list.

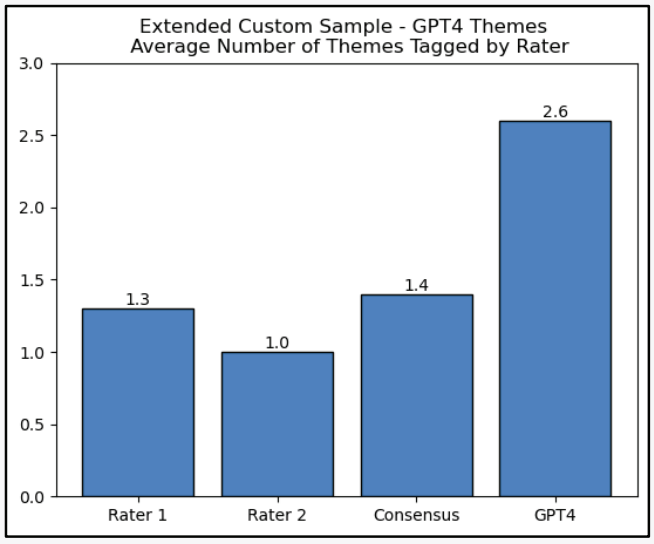
This process yielded a list of nine refined themes and definitions that were representative of the comments across all three samples, as well as three themes that emerged as uniquely representative within each individual sample. Because these themes were derived from those that originally emerged from our samples, we refer to the human-refined themes as "superordinate". Our subsequent analyses considered our Extended Custom Sample in the light of 10 total superordinate themes (nine that overlapped conceptually with the other samples, and one that was unique). (See Appendix, Table 5.)

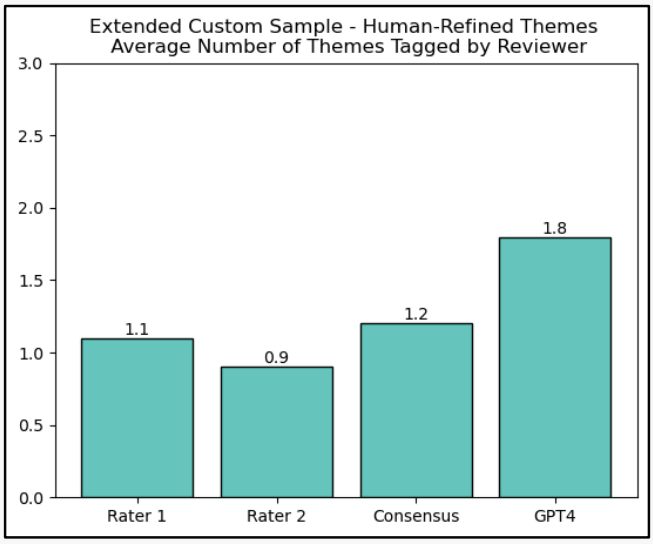
Rather than repeating the time-intensive human coding task, we mapped each of the superordinate themes to the appropriate original themes tagged. This gave us a pseudo-human-tagged Extended Custom Sample, which we then compared with a new GPT4 output created by repeating our multi-theme tagging process. This time, we included the new human-refined superordinate themes and definitions as part of the input for GPT4's task.

Results

Conducting the same comparisons, we discovered that GPT4 tagged an average of 1.8 superordinate themes per comment, down from the 2.6 themes originally tagged. Additionally, our pseudo-human consensus yielded 1.2 superordinate themes per comment, down from 1.4.

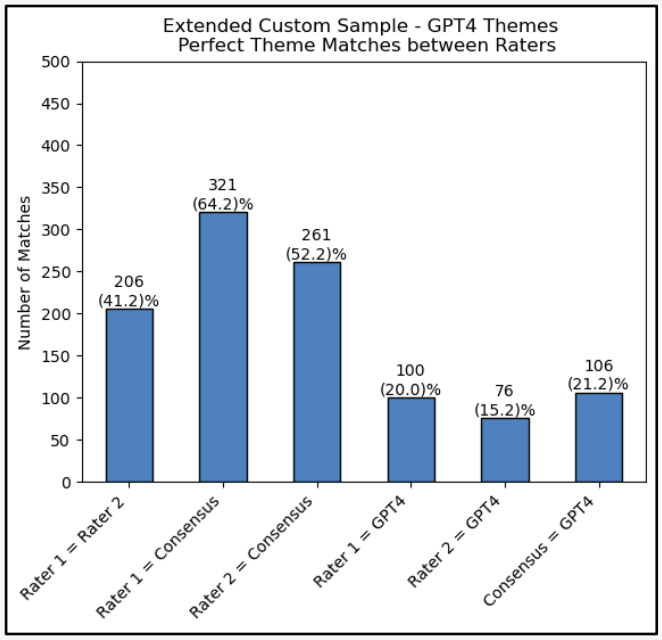
AVERAGE NUMBER COMPARISONS

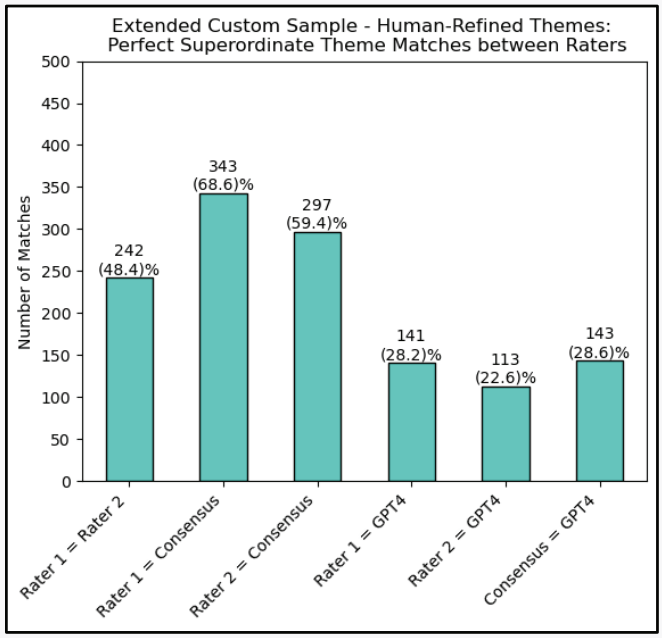




The list of superordinate themes provided by GPT4 matched perfectly with the consensus coder mapping 28.6% of the time (κ = 0.255), which was up from 21.2% of the time on the original array of themes (κ = 0.126).

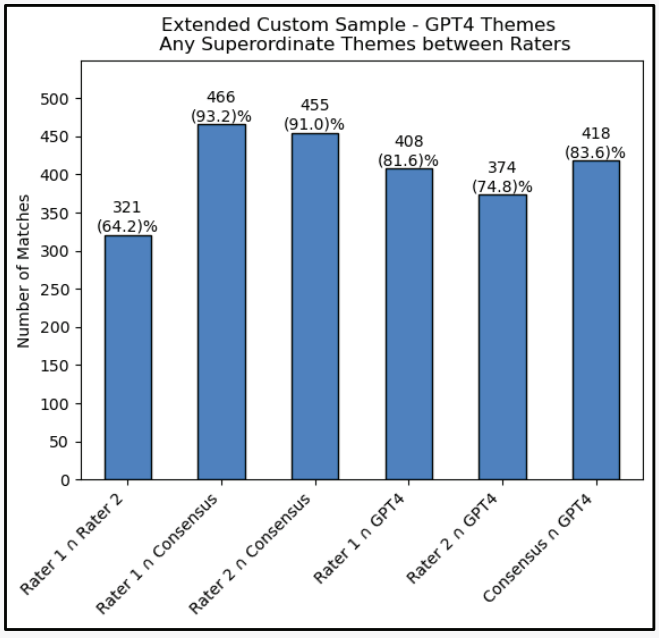
PERFECT MATCH COMPARISONS

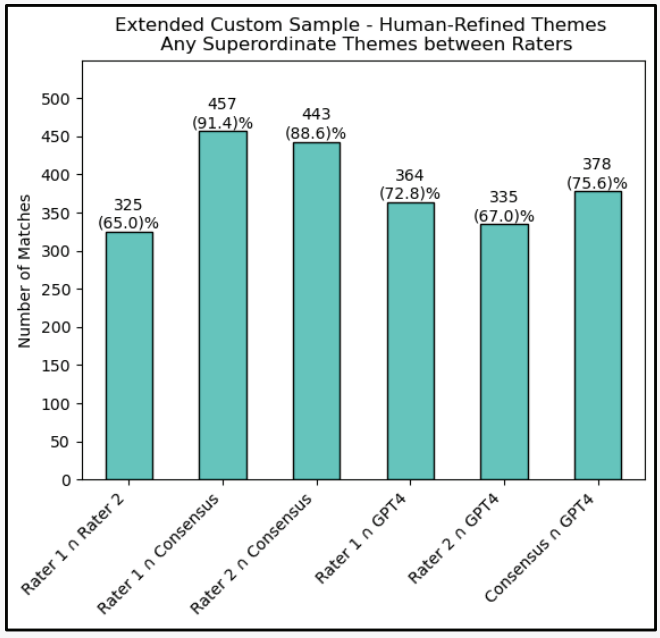




The list of tagged superordinate themes provided by GPT4 and the list of mapped superordinate themes from consensus had at least one common theme 75.6% of the time. This was down from the previous 83.6%, but still relatively high.

SUPERORDINATE THEMES COMPARISONS





The increase in agreement between human coders and slight decrease between the consensus and GPT4 seems to indicate the human-reviewed superordinate themes provided a more distinct framework to work from. Because of our human intervention, GPT4 tagged fewer themes overall while still identifying at least one appropriate theme a majority of the time. Because the themes were more distinct, we suspect the co-occurrences observed in this iteration were closer to a true correlation between the themes found within participant's comments than in the previous one.

Lessons Learned

Based on our observations at the time of this study, we consider GPT4 an impressive supplementary tool for text-based thematic coding rather than a viable replacement for human coders. It can identify themes from open-response comments, though those themes likely may not be as distinct and meaningful as human-identified themes. It can also tag themes to large collections of comments at similarly reliable rates to human coders. However, it tends to be more accurate when given the option to over-tag rather than discern the *best* theme. Finally, there's no question that GPT4 is much faster than human coders for this task. The average time for GPT4 to tag a round of 500 comments was less than an hour per sample, where for our human coders, this task took between two and four hours per sample.

As is typical for AI-based tools, the trade-off seems to be in efficiency (AI's strength) or nuance (still a human strength, at least in our case). We provide the following suggestions to help strike a balance between the two when engaging with publicly accessible AI tools, turning that *or* into an *and*.

1. **Define Your Themes**
   Perhaps the biggest takeaway from this study is to spend more time defining themes before engaging GPT4 to tag them. In the case of this study, spending some time refining the themes GPT4 recommended led to it having a stronger ability to tag and interpret those themes. For more specialized datasets or in cases where certain themes are anticipated, defining a thematic structure before employing GPT4 for the tagging activity could help with leveraging the efficiency of GPT4 while retaining some of the nuance of human expertise.
2. **Treat AI Like One Coder Among Many**
   Just like only using one human coder for qualitative data analysis, relying exclusively on a tool like GPT4 may produce biased or inaccurate results. However, due to the ease and speed with which GPT4 can classify open-response comments into themes and produce similar results to that of a human coder, it could serve as a great additional coder, preliminary explorer, or summarizer in cases where more rigorous methods may not be necessary. We recommend in those cases always checking its work, including paying close attention to situations where it may have over-tagged themes without discretion.
3. **Pair Qualitative Insights with Quantitative Insights**
   Although we focused entirely on text-based insights for this study, in an evaluation setting, it often makes a lot more sense to consider the results in the context of the other data collected. If your plans are to integrate AI into analytics and reporting processes, we recommend pairing the identified themes with the context provided by other collected data. This context will help with identifying whether the AI model might have missed the mark in its identification of themes.

Our ultimate goal for this study was to better understand the impact integrated coaching had on leaders' development while using AI to assist in large-scale text analysis. The blending of efforts between AI and human reviewers led to insights that we shared in an online companion piece to this paper. We share the final, refined themes and their prevalences across samples below (See Appendix, Table 6).

Limitations and Conclusion

In this study, we engaged with OpenAI's GPT4 to explore 1,500 participant responses to an open-ended item appearing across several feedback surveys associated with coaching in CCL's leadership development programs. By leveraging GPT4, we were able to better understand its usefulness and limitations for this task, while also gaining insight into the unique impact coaching can have in leadership development programs from the perspective of participants.

However, our results come with several limitations and considerations for those who may be thinking of using AI in their own data analysis tasks. First, although we anticipated many of the potential drawbacks of engaging an AI model in this task – hallucinations, biases, and lack of specificity to the leadership development context our data comes from – we did not recognize that the model we were using had a character limitation of 38,000 characters. Because the data input we used to first generate themes for each sample included 500 comments and anywhere between 50,000 and 60,000 characters, this likely influenced GPT4's ability to holistically summarize the initial themes we used for the study. Adhering more closely to the documented limitations of the model may have led to more accurate results, and we suggest doing so when leveraging AI to summarize large amounts of text data.

As well, in our study, we refrained from engaging in more sophisticated techniques for using AI, like fine-tuning the model for our linguistic context. Although this was an intentional step to reflect a use of AI that is more accessible to a wider range of potential users, we acknowledge that more sophisticated processes exist that would have likely led to more specificity in our AI-generated results and perhaps less need for human intervention.

Finally, we acknowledge there are some key differences between the methods we followed in our study and those someone might engage in using a web-based AI interface (like ChatGPT, Claude, DeepSeek, or the many others currently available). In general, the process for appending input (like a text-based dataset) to a prompt differs from the API-based process depending on the interface and model used. In fact, it may be illegal or ill-advised to do so in cases where the data in question are sensitive, proprietary, or otherwise should not be shared.

On a more technical note, we noticed that when we accessed the API in a single session, the results it returned were likely to be influenced by any previous interaction. Because of that, in the process of tagging themes to 500 separate comments, our first round resulted in GPT4 hallucinating the same theme over and over, even for comments that were unrelated to that theme. We avoided this by initiating a new session for each theme tagging task (so, 500 sessions per sample); however, this is not practical in an environment where someone may be using a web-based interface. It also does not reflect the safeguards that may be built into a web-based interface to keep this from happening in that environment. For those planning to try this out using a web-based AI interface, we suggest being cautious of the possibility of this type of hallucination.

As AI evolves in its ability to assist with day-to-day tasks, including data analysis tasks, we encourage both researchers and casual AI users alike to continue to test and share new modes of human-AI collaboration. Our study reflects a need to better understand the sophistication with which these tools produce not only accurate but meaningful output in a data analytics context. A world where efficient data insights are at our fingertips is enticing, yet human guidance remains necessary to ensure those insights contain nuance rather than noise.

References

Bukin, S., & Schneider, M. (2022, March 17). *Want greater impact? 7 steps to center your customers' voices.* Center for Creative Leadership Research and Innovation Site. https://cclinnovation.org/news-posts/want-greater-impact-7-steps-to-center-your-customers-voices/

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

Khan, A. H., Kegalle, H., D'Silva, R., Watt, N., Whelan-Shamy, D., Ghahremanlou, L., & Magee, L. (2024). Automating thematic analysis: How LLMs analyze controversial topics. *Microsoft Journal for Applied Research, 21,* 69-87. https://doi.org/10.48550/arXiv.2405.06919

McHugh, M. L. (2012). Intercoder reliability: the kappa statistic. *Biochemia medica, 22*(3), 276-282. https://hrcak.srce.hr/89395

Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research, 15(1), 45-55.

About the Authors

Katelyn McCoy

Katelyn McCoy is a senior research scientist on CCL's Insights & Impact team. She oversees many aspects of data collection and reporting for CCL's leadership development offerings. She also leads customized client evaluation studies, particularly for coaching initiatives, and contributes to key research projects related to the future of leadership.

Tanner Landolt

Tanner Landolt is a senior data analyst at NuView Analytics and previously worked with CCL as a data analyst. With a background in industrial organizational psychology and applied data analytics, Tanner applied his expertise to help modernize CCL's executive coaching reporting processes and uncovered new ways to summarize and share text-based data insights.

Ezekiel Welsh

Ezekiel Welsh is an evaluation analyst with experience in employee data and leadership development program evaluation. He works primarily with CCL's Insights & Impact team, collaborating on research and studies of leadership development impact, managing client dashboards, and investigating innovative ways to share qualitative insights with CCL's clients.

Caroline DeStefano

Caroline DeStefano is an implementation manager responsible for executing the end-to-end delivery of CCL's high-impact leadership development programs globally. She is also a certified coach and founder of The Leadership Side, where she offers coaching and tailored leadership development solutions. Her partnership and expertise were crucial to ensure our analysis of leaders' comments remained grounded in the program participant point of view.

Appendix

TABLE 1. DESCRIPTIVE STATISTICS – SURVEY RESPONSES

Statistic	Open Enrollment Survey	Custom Survey	Extended Custom Survey
Response Count	1,069	4,070	962
Average Word Count	20.47	17.79	21.09
Median Word Count	16	15	17
Standard Deviation Word Count	14.69	13.57	17.08
Total Character Count	124,648	412,249	115,317
Average Character Count	116.58	101.04	119.87
Median Character Count	95	84	98
Standard Deviation Character Count	82.66	75.84	93.88

Statistic	Open Enrollment Sample	Custom Sample	Extended Custom Sample
Response Count	500	500	500
Average Word Count	20.90	17.58	21.40
Median Word Count	17	14	17
Standard Deviation Word Count	14.76	13.71	17.15
Total Character Count	59,628	49,773	60,742
Average Character Count	119.26	99.55	121.48
Median Character Count	96	83	98.5
Standard Deviation Character Count	83.22	76.75	92.98

TABLE 2. GPT4-GENERATED THEMES BY SAMPLE

Open Enrollment Sample	
Self-Awareness	Many participants gained new insights about themselves, furthering their understanding of their own leadership style, strengths, and weaknesses.
Goal Setting	Individuals felt that coaching enabled them to set clear, actionable goals for their leadership development.
Validation	Coaching validated participants' self-perceptions and experiences, affirming their instincts and existing knowledge.
Understanding of Assessments	Participants frequently mentioned that coaching helped them to better understand their assessment results and how to interpret them in context.
Improvement Strategies	Coaching provided specific strategies and techniques to improve leadership ability, manage challenges and leverage strengths.
Connection & Context	Many participants felt that coaching helped to link their professional experiences, feedback from their 360-degree assessments, and personal goals, providing a powerful context for their development.
Increased Confidence	Coaching resulted in increased self-confidence about their ability to face challenges and make necessary changes in their leadership style.
Perspective Shift	Coaching helped participants to consider different perspectives, challenging their beliefs and encouraging them to consider alternative approaches.
Personalized Attention	Coaching offered a unique, personalized experience that was seen as highly valuable by the participants.
Encouragement	Coaches offered support, motivation, and reassurances which helped participants build confidence and resilience.
Useful Problem-Solving Techniques	Participants learned or refined problem-solving techniques, leading to increased leadership effectiveness.
Behavior Adjustments	Participants identified specific behaviors to change or enhance to improve their leadership effectiveness.
Focus on Actionable Steps	Coaches helped participants distill information into actionable steps, making it easier to implement changes.
Enhanced Communication Skills	Participants learned to communicate better with their colleagues and subordinates, thereby improving their leadership efficacy.
Carrying Lessons Back to the Workplace	A key benefit for many was the ability to apply learnings from the coaching directly to their roles back at work.
Custom Sample	
Increased Self-Awareness	Many leaders mentioned gaining insights into their strengths, weaknesses, behavior patterns, and how they are perceived by others.
Goal Setting and Planning	Coaching helped in clarifying, setting, and organizing personal and professional goals, along with actionable steps to achieve them.
Improve Communication Skills	Many mentioned that the coaching helped them better understand and improve their communication towards their team and peers.
Increased Perspective and Reflection	Coaching brought out different perspectives, increased self-reflection and made leaders more mindful about their attitude and approach to issues.
Understanding and Applying Feedback	Leaders appreciated the help in interpreting and applying feedback from assessments, especially the 360-degree feedback, as well as identifying development areas.
Improved Confidence & Empowerment	Several leaders saw an increase in their confidence and empowerment, becoming more open to acknowledging their accomplishments and more self-assured in their abilities.
Handling Challenges	Coaching helped leaders identify and address their key leadership challenges, providing potential strategies to overcome them.

TABLE 2. GPT4-GENERATED THEMES BY SAMPLE, CONTINUED

Enhancing Emotional Intelligence	There was growth in understanding, managing, and channeling emotions productively.
Improved Listening Skills	Leaders recognized the importance of active listening and changing the way they approach conversations.
Bridging Learning & Application	Leaders valued coaching as a link between what they learned in theory and how to apply it in their leadership roles.
Importance of Balance	Some leaders mentioned coaching helped them to see the importance of balance between work and home life.
Support and Encouragement	Leaders appreciated the support, encouragement, and non-judgmental environment provided by the coaches.
Enhancing Leadership Style	Leaders realized the importance of improving their leadership style, including increasing empathy, accepting failure, and inspiring others.
Providing Tools	Many leaders appreciated coaches' provision of concrete tools, methods, and resources to implement positive change.
Improvement in Self-Care	Some leaders indicated they got advice to take better care of themselves both mentally and physically.
Extended Custom Sample	
Self-understanding	Coaches were instrumental in helping leaders understand their own leadership style, strengths, weaknesses, behaviors, and values.
Strategy & Action Planning	The coaching facilitated the creation of clear and actionable plans for growth and development.
Improved Confidence	The coaching sessions boosted confidence levels in leaders and helped them see their full potential.
Perspective & Insight	Coaches offered fresh, objective viewpoints that helped leaders see things in a new light, and helped them internalize assessments and feedback.
Focused Development	Coaches helped leaders identify and prioritize key areas for improvement and growth.
Application of Learning	The coaching aided in connecting the dots between training materials, discussions, assessments, and the leaders' day-to-day roles.
Enhanced Communication	Coaching assisted leaders in developing and improving their communication with their teams and superiors.
Accountability	The coaching process provided an element of accountability for leaders to follow through on the action steps identified.
Empathy and Supportive Environment	Coaches were appreciated for their empathetic listening and for creating a safe space for open conversation.
Increased Self-Awareness	Leaders gained insights into their blind spots and tendencies that were affecting their leadership style.
Practical Suggestions	The advice and tools provided during the coaching sessions were viewed as useful and applicable to everyday situations.
Integrative Approach	Coaches were appreciated for bringing together various assessment results and leadership challenges into understandable frameworks for the leaders.
Motivation to Improve	The coaching experience motivated leaders to embrace their development areas and work towards improvement.
Tangible Impact	Leaders felt coaching had a real and tangible effect on improving their leadership skills.
The Value of Time	Having dedicated time and space to think and reflect was recognized as very valuable for leaders.

TABLE 3. INSTANCES WHERE CONSENSUS CODERS WERE MORE ACCURATE THAN GPT4

Response	GPT4	Consensus
He was very kind and easy to talk to and I felt like he had great perspectives. I felt like he really listened when I talked. I think he made want to speak with an executive coach more often.	Improved Listening Skills	Increased Perspective and Reflection
She was quick to identify my shortcoming and challenged me to move forward with positive reinforcement and creative thinking.	Handling Challenges	Support and Encouragement
Very good experience, it helped me a lot in my self-knowledge and generation of action plans.	Increased Self-Awareness	Goal Setting and Planning

TABLE 4. INSTANCES WHERE CONSENSUS CODERS CORRECTLY DISTINGUISHED "NON-THEMES" WHEN GPT4 RETURNED A THEME

Response	GPT4	Consensus
Made me look at the less positive points.	Increased Self-Awareness	I don't know
Helpful to focus on the critical elements.	Goal Setting and Planning	I don't know
It was the highlight of the program and I look forward to learning more from her.	Encouragement	I don't know

TABLE 5. HUMAN-REVIEWED THEMES AND DEFINITIONS

Theme	Sample Appearance	Description
Self-Awareness	All Samples	Leader gained insight into their own strengths, weaknesses, and behavior patterns and how they are perceived by others.
Perspective	All Samples	Coach helped the leader see new, objective viewpoints on their situation or challenges.
Goal Setting & Action Planning	All Samples	Leader was able to set personal and professional goals and identify specific actionable steps to achieve them, including behaviors to change.
Sensemaking Around Feedback and Assessments	All Samples	Coach helped the leader to make sense of their assessment results and feedback to identify ways to improve their leadership.
Self-Efficacy	All Samples	Coaching increased the leader's confidence, made them feel more empowered, more open to acknowledging their own accomplishments, more self-assured in their abilities, and motivated to continue developing as a leader.
Application of Learning	All Samples	Coach helped the leader to understand the connection between training materials and their day-to-day roles, leading to an enhanced ability to apply what they learned.
Support	All Samples	Coach provided the leader with a supportive, encouraging, and non-judgmental environment.
Communication	All Samples	Leader improved their communication with teammates, subordinates, peers, and superiors, through better listening and speaking techniques.
Providing Tools or Solutions	All Samples	Coach provided specific strategies, tools, methods, or resources to help the leader overcome their unique challenges.
Personalization (Unique Theme)	Open Enrollment Sample	Coaching offered a unique opportunity for leaders to make their learning personal, beyond other learning activities.
Goal Progress (Unique Theme)	Extended Custom Sample	Coach held the leader accountable to follow through on the goals or action steps identified.
Wellbeing (Unique Theme)	Custom Sample	Coach helped the leader to see the importance of work/life balance and self-care for enhancing their ability to lead effectively.

Note: When refining the themes, Personalization, Goal Progress, and Wellbeing were not common enough across samples to fit into the overall thematic structure. We specified them as unique themes as they appeared to hold unique meaning within their sample.

TABLE 6. PREVALENCE OF HUMAN-REFINED THEMES ACROSS SAMPLES

Theme	Open Enrollment Sample	Custom Sample	Extended Custom Sample
Perspective	**13%**	**17%**	**26%**
Self-Awareness	**15%**	**13%**	**14%**
Sensemaking Around Feedback and Assessments	**15%**	**13%**	4%
Goal Setting & Action Planning	6%	10%	**19%**
Application of Learning	**13%**	5%	7%
Personalization (Unique to Sample)	**13%**	-	-
Wellbeing (Unique to Sample)	-	3%	-
Goal Progress (Unique to Sample)	-	-	4%
Providing Tools or Solutions	9%	7%	6%
Support	3%	4%	5%
Self-Efficacy	1%	2%	4%
Communication	1%	2%	2%
Theme - Other	1%	16%	0%
I don't know	10%	10%	9%

Note: This table contains the percentage of comments in each of our samples to which our refined themes were tagged by human coders. As a general rule, we considered themes with a percentage higher than 10% to be "prevalent" in any given sample, and we considered themes with a prevalence rate in one sample that was at least one standard deviation higher than the rest to be "uniquely prevalent". Both "prevalent" and "uniquely prevalent" themes are bolded. "Unique to Sample" themes emerged from their samples but were not reflected in the others.

Multiple Theme Tagging Prompt (Exact Wording)

I'm looking at a response to this prompt: Please comment how coaching impacted your leadership development experience.

Below is a comment from a leader who experienced coaching as part of a leadership development program, as well as a list of themes.

Return the theme or themes from the list that best categorize the comment, separated by a comma.

Do not return the theme definition.

Only return the themes like they are listed with no added punctuation.

Do not hallucinate a different theme.

If you do not know, return "I don't know".

Comment:

CCL LOCATIONS

Americas

+1 336 545 2810

ccl.org

Europe, Middle East, Africa

+32 (0) 2 679 09 10

ccl.org/emea

Asia Pacific

+65 6854 6000

ccl.org/apac

Greater China

+86 21 6881 6683

ccl.org/china

Center for Creative Leadership

The Center for Creative Leadership (CCL)® is a top-ranked, global, nonprofit provider of leadership development. Over the past 50 years, we've worked with organizations of all sizes from around the world, including more than 2/3 of the Fortune 1000. Our cutting-edge solutions are steeped in extensive research and our work with hundreds of thousands of leaders at all levels.