Fall 2023

Evaluating AI Sentiment Analysis

Aakriti Shah
u2aakriti@gmail.com

EVALUATING AI SENTIMENT ANALYSIS

Aakriti Shah
Rollins College
ashah@rollins.edu

ABSTRACT

This paper presents a comparative analysis of human and AI performance on a sentiment analysis task involving the coding of qualitative data from community program transcripts. The results demonstrate promising but imperfect agreement between two AI models, Claude and Bing, versus three human annotators and one expert annotator using the Community Capitals framework categories. While both models achieved fair alignment with human judgment, confusion patterns emerged involving metaphorical language and text overlapping multiple categories. The findings provide a case study for benchmarking conversational AI systems against human baselines to reveal limitations and target improvements. Key gaps center around distinguishing between social and human elements and handling cultural references. Expanded testing on more diverse datasets could further quantify differences in classification capabilities. Overall, the analysis exposes definable areas where machines still struggle compared to humans, highlighting productive research directions to eventually achieve a similar threshold to humans across diverse language inputs. As AI systems enter real-world applications, human-AI comparative studies can help define boundaries between robust statistics-based learning and adaptive human cognition.

***K*eywords** sentiment analysis · AI models · large language models (LLMs) · Ripple Effect Mapping (REM) · thematic analysis · kappa statistics · natural language processing (NLP) · qualitative analysis · categorical sorting · labeling data · Community Capitals · human-centered computing

1 Introduction

In recent years, the application of artificial intelligence (AI) to tasks traditionally performed by humans has expanded rapidly. Natural language processing (NLP), computer vision, speech recognition, and other AI technologies now rival or surpass human capabilities on certain benchmark datasets [1]. However, the accuracies of AI systems on focused tasks do not capture the full spectrum of human cognition. While humans have the ability to integrate context, common sense, emotional reasoning, and even sarcasm in adaptable ways, AI systems have trouble with these tasks. Evaluating AI on real-world unstructured data requires comparison to human performance. This type of human-AI comparative study could reveal the limitations of current systems and provide insight into the development of next-generation capabilities [2].

One active area of NLP research is thematic analysis, which involves "identifying, analyzing, and reporting patterns (themes) within data" [3]. Large language models (LLMs), such as Claude by *Anthropic* and Bing by *Microsoft*, have shown promising performance on thematic analysis tasks. However, systematic comparisons to human performance are still needed to understand their limitations. This paper presents an analysis comparing human coding and AI coding of qualitative data from a community program evaluation. The results provide the strengths and weaknesses of when AI systems are applied to a sentiment analysis task.

The data was sourced from group discussions run as part of an impact assessment for *Crave of Central Florida* [4], a leadership development program for social innovators and entrepreneurs in the Orlando area that focuses on personal and community development. The transcripts of the interview sessions from *Crave* were categorized and

[1]This project was conducted collaboratively with James McIntyre and James Temple, as co-contributors to the content of this paper.

coded using the Community Capitals Framework, which categorizes statements into one of 8 categories: *Natural*, *Cultural*, *Human*, *Social*, *Political*, *Financial*, *Built*, and *None* Capitals. This paper compares the coding of Crave discussion transcripts by three human annotators, an expert annotator, and two AI models, *Claude (Anthropic)* and *Bing (Microsoft)*. AI developers can utilize such comparative studies to better understand gaps between human and machine cognition that should be addressed in future research. This research could also simultaneously help Crave understand the themes that are being referred to the most during these *Ripple Effect Mapping* (REM) sessions.

2 Literature Review

2.1 Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding

Supporting Qualitative Analysis with Large Language Models explores the potential for leveraging recent advances in large pre-trained language models to assist with qualitative analysis tasks like sentiment coding. According to Xiao et al. [5], by combining expert-curated Codebooks with flexible prompting of models like GPT-3, the authors achieve promising performance in deductive coding with only a model pre-trained on general data. This indicates the power of transfer learning for sentiment classification without extensive domain-specific fine-tuning. As I evaluate latest methods in AI sentiment analysis, this provides valuable evidence for the potential of large foundation models to achieve fair or even substantial agreement with human raters without overfitting to narrow datasets. The agile prompting approach may point toward more generalize solutions in the future.

2.2 LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding

Chew et al. [6] delivers an extensive empirical benchmark leveraging large language models for deductive coding across diverse textual datasets. By prompting models like GPT-3 in a format familiar from NLP data rather than expecting output to match human annotations, they enable scaled assistance for qualitative coding with accuracy rivaling expert human raters. This indicates the power of transfer learning from models trained on broad data to streamline tasks like sentiment analysis without losing quality. The efficiency gains also showcase the data processing advantages of AI over manual approaches given appropriate evaluation. As I assess merits of modern AI sentiment systems versus human evaluation, both accuracy and speed are essential considerations, so this work provides a relevant template for rigorous AI benchmarking against human performance.

2.3 Sentiment Analysis on Twitter Data using Apache Spark Framework

The Sentiment Analysis on Twitter Data Using Apache Spark [7] examines real-world application of AI sentiment classification at large scale. While highlighting expansive data processing capacity exceeding human limitation, examining model shortcomings like sarcasm detection also emphasizes current gaps I must consider when evaluating state-of-the-art systems against human raters. This serves to demonstrate not only promising potential but also existing challenges in deploying AI for nuanced sentiment analysis across diverse informal textual data. My assessment therefore scrutinizes how well AI models can capture complex emotions that might need an amalgamation of human and AI input to analyze accurately.

2.4 Large Language Models Understand and Can Be Enhanced by Emotional Stimuli

This paper provides an insightful framework for evaluating whether large language models (LLMs) can genuinely comprehend emotional cues and stimuli like humans. Such emotional intelligence gives people distinct advantages in complex problem-solving. The authors conduct comprehensive experiments across 45 diverse tasks, testing leading LLMs including GPT-4. Li et al. [8] demonstrate that LLMs do exhibit a grasp of emotional intelligence, as performance improves when augmenting prompts with psychological phrases (dubbed "EmotionPrompts"). For example, accuracy relatively increased 8% on instruction induction and over 100% on BIG-Bench tasks. A human study of over 100 participants further confirmed EmotionPrompts boost performance on generative tasks by nearly 11% on metrics of quality, truthfulness and responsibility.

The paper discusses factors driving this efficacy, like EmotionPrompts better activating inherent reasoning abilities. But deficiencies handling atypical nuanced cases parallel findings in medical imaging assessments, where despite efficient data processing, algorithms currently lack humans' versatile perception honed through grounded experience. This indicates exciting potential for hybrid AI that synthesizes statistical learning with empathy modeling. Overall, these emotional intelligence experiments pioneer interdisciplinary methods to qualitatively evaluate LLMs against human cognition. Demonstrating performance gains from emotional appeals could inform techniques for

improving sophisticated tasks dependent on contextual reasoning. But limitations on convoluted inputs highlight the enduring need for versatility that is reminiscent of human intelligence.

3 Methods

3.1 Data Collection

The text data utilized for this analysis was sourced from transcripts of group discussions and *Ripple Effect Mapping* (REM) sessions conducted as part of an impact assessment for Crave of Central Florida. Crave is a nonprofit leadership development program that serves young professionals and social entrepreneurs in the Orlando area who are engaged with issues of spirituality and social entrepreneurship. The program provides training on topics such as communication and spiritual development in a small group setting. Crave aims to support participants' development as leaders who can effectively contribute to Central Florida's nonprofit ecosystem.

The Impact Lab research team from Rollins College conducted two REM sessions with Crave stakeholders in 2021 [9]. REM is a qualitative evaluation method that elicits participants' perspectives through structured group discussion and collaborative mind mapping. The sessions generated transcripts of dialogue between Crave participants, alumni, staff, and facilitators. For this study, a subset of 107 statements were extracted from the Crave REM transcripts. The statements covered topics related to Community Capitals, including culture, society, politics, economics, and spirituality in Central Florida.

3.2 Annotation Process

Three undergraduate computer science students (student annotators) and a computer science professor (expert annotator) independently coded each of the 107 statements using the Community Capitals Framework. This framework defines seven types of community assets: Natural, Cultural, Human, Social, Political, Financial, and Built Capital. A code of "None" was also permitted if a statement did not pertain to any category.

Capital	Description
Natural (N)	The natural resources and environment of a particular place, which can include traditional exploitable resources, but also encompasses weather, natural features, beauty, and geographic location.
Cultural (C)	The traditions, language, heritage, and other expressions of the peoples that live in a community.
Human (H)	The ability for community members to acquire new resources, develop knowledge, and pursue opportunities. This capital also encompasses the values of inclusive and proactive community leadership.
Social (S)	Connections between individuals and organizations that contribute to development. Bonding Social Capitals are tighter connections that build cohesion within a community. Bridging Social Capitals are looser ties that connect different organizations and subgroups.
Political (P)	Access to centers of power, government officials, and the political process. This capital also encompasses the capacity of community members to advocate for their own interests.
Financial (F)	Assets available to support capacity-building, including investments in business, social entrepreneurship, and future community development.
Built (B)	Physical infrastructure that supports the other capitals, including buildings, transportation, utilities, and digital connectivity.
None (X)	"None": the item does not clearly relate to any of the Capitals.

Table 1: Capitals Framework [10]

The annotation process followed these steps:

The annotators were provided with the Community Capitals Framework category definitions and instructed to do the coding task.

1. The 3 undergraduate annotators and the expert annotator were provided with the Community Capitals Framework and the category definitions.
2. Each annotator independently labeled each of the 107 statements from the Crave interview sessions with exactly one of the framework capital codes, or "None".
3. The 3 students' individual labels for each statement were compared. The student annotators then manually went through the statements and their associated labels to come to a consensus on what the appropriate label was for each statement. This label was set as the 'ground truth', or 'human modal' column.
4. Each annotator provided the Claude and Bing AI systems with the same prompt – 107 statements, Community Capitals Framework, and instructions on how to code each statement. For each AI system, 2 student annotators did 2 runs, while the final student annotator did 3 runs. This was done to efficiently analyze the data later on and determine each annotator's self-consistency.

The systems independently generated predicted labels for each statement, producing 7 runs from the Claude AI and 7 from the Bing AI. In total, this produced 18 runs of data for each statement. In addition to this, there were also 3 modal runs for the humans (ground truth), Claude AI (Claude Modal), and Bing AI (Bing Modal). This process produced a truth benchmark dataset based on human consensus, as well as modal predicted labels from the AI systems for quantitative comparison to the truth.

Recent advancements in AI systems have led to the development of Claude by *Anthropic* and the Creative Mode for Bing AI by *Microsoft*. Comparing these two systems allows for a more insightful exploration of their respective capabilities for sentiment analysis because of how they contrast. Bing has the benefit of drawing from the vast data and research resources available to a major technology company like *Microsoft*. In contrast, Claude was created by the startup *Anthropic*, using a more focused approach centered around constitutional AI methods intended to improve safety. Benchmarking Claude versus Bing Creative Mode provides quantification of performance differences between an AI system produced with ample corporate resources and another that is reliant on algorithmic innovations. Examining empirical results between the two models, side-by-side, reveals differences in how underlying training processes manifest in actual conversational interactions.

4 Results

4.1 Quantifying Qualitative Data

The level of agreement between the data of each rater (with multiple runs – human, Claude AI, and Bing AI) was quantified using Fleiss' kappa. The expert annotator was not included in this kappa analysis because Fleiss' kappa requires at least 2 columns of data, and the expert annotations only consisted of one column of codes. Fleiss' kappa is a statistical measure that is used to quantify the agreement between multiple runs of the same rater on assigning categorical ratings, indicating self-consistency. For example, it can measure how effectively the different Claude runs agree with each other in assigning ratings. Similarly, it can determine the agreement between the different Bing classification runs. Additionally, the level of agreement between two different raters was visualized by creating a confusion matrix, with *Python* code, to determine the percentage of labels that matched between both raters. Cohen's kappa statistic is a metric that is used to compare the way that two raters coded each statement against each other (Claude vs. Human, Bing vs. Human, Claude vs. Bing, Human vs. Expert, Expert vs. Claude, and Expert vs. Bing), also known as inter-rater consistency. Both of these metrics, Cohen's and Fleiss' kappas, serves as a way to determine the consistency between two raters that goes a step further than simply calculating the accuracy; this accounts for by-chance agreement. The scores for both Cohen's and Fleiss' kappas range from 0 – chance-level agreement – to 1.0 – perfect agreement. The modal data from the 7 Claude runs and the 7 Bing runs were compared against the truth labels using Cohen's kappa.

The confusion matrices were computed between each pair of raters and the overall accuracy and kappa scores are reported in the Results section. This allows for the evaluation of the consistency of each AI's predictions across runs and how their performance on the annotation task compares to the human annotators. The utilization of these kappa statistics allowed for conducting a systematic, impartial, and quantitative evaluation of self-consistency and inter-rater consistency, considering all essential criteria simultaneously.

McHugh provides guidelines for interpreting these kappa values:

Kappa Value	Agreement Level
≤ 0	No Agreement
0.01 - 0.20	Slight Agreement
0.21 - 0.40	Fair Agreement
0.41 to 0.60	Moderate Agreement
0.61 to 0.80	Substantial Agreement
0.81 to 0.99	Near Perfect Agreement
1	Perfect Agreement

Table 2: McHugh's Kappa Agreement Guidelines [11]

The key results were found through Python script and spreadsheet calculations and are demonstrated in the tables below:

Rater	Fleiss' Kappa
Human Annotators	0.31
Claude AI	0.24
Bing AI	0.45

(a) Fleiss' Kappa Scores

Rater vs. Rater	Cohen's Kappa
Human vs. Claude	0.38
Human vs. Bing	0.35
Bing vs. Claude	0.50
Expert vs. Human	0.37
Expert vs. Claude	0.33
Expert vs. Bing	0.37

(b) Cohen's Kappa Scores

Figure 1: Kappa Scores

5 Discussion

The self-agreement of the Human Annotators with a Fleiss' kappa score of 0.31 indicates a fair level of consensus between the human coders. While imperfect, this provides a reasonable benchmark for evaluating the AI models' performances. The Cohen's kappa values of both of the AI models against the Human Annotators are in the fair range, Bing (0.3476) and Claude (0.3826); this demonstrates promising but imperfect agreement between the AI models and the undergraduate students. Claude performs closer to the human consensus than Bing does, which indicates closer alignment of its predictions with human judgment. However, both models failed to match human codes over 40% of the time, highlighting ample room for improvement.

Claude and Bing AI performed well classifying Natural Capital examples but struggled with the more prevalent and critical Social and Human Capital classifications. Both models achieved just 50 - 55% accuracy on these vital societal building blocks, undermining reliability for real-world application. Agreement between the models was also mixed – while perfectly aligned on trivial Natural examples, Claude and Bing AI displayed concerning differences while categorizing essential Human Capital statements. As Human Capital was highly frequent in the underlying data, inability to consistently encode such a key element diminishes credibility across both systems. Ultimately, Claude and Bing AI's stellar Natural Capital performance is overshadowed by critical weaknesses in their inability to accurately classify Societal and Human statements. Targeting improvement in these central competencies is imperative to increase accuracy for both models. Overall, the models aligned perfectly only on insignificant examples like Natural and Built, while demonstrating divergence in categorizing crucial Capitals like Human and Social.

Claude and Bing agreed with each other more often than they agreed with humans. This reveals limitations in both models' ability to match the manner in which humans go about annotating. It should also be noted that the AI

models agreed more with the 3 undergraduate annotators than they did with the expert annotator. It is possible that these AI models could require more context so that it is able to understand the task fully. Also, the sample size of 107 statements is small, so it only provides a preliminary guide for this investigation. The "fair" Fleiss' kappa score between the annotators could also contribute to the lower Cohen's kappa scores for the Humans Annotators vs. AI Models and Expert Annotator vs. AI Models. While analyzing the coding results, more common misclassification patterns were found throughout the data after having done the kappa statistics analyses. These patterns are described in subsections 5.1 - 5.5.

5.1 Social vs. Human Capital

Consistent classification confusion arose between the similar Social and Human Capital codes. These socially-focused codes carry the greatest weights in analysis due to their high frequencies in the underlying annotations. However, clear demarcation between social and human elements depends greatly on interpretation of context. Even human annotators demonstrated disagreement categorizing these pivotal codes, reflected in only fair Cohen's kappa agreement levels. As such, while mixing Social and Human classifications certainly contributed to lower AI accuracies against human judgment, even humans failed to fully align on these prominent codes. As the core contributors to total capital calculations, improving the ambiguity between these capitals should remain a priority for advancing reliability. Some examples of where the AI models could not determine whether a statement is Social or Human Capital are as follows:

Statement	Rater	Rater Code	Agreed Code
"You don't have to present or code switch; you get to the real and get better at the real"	Bing	H	S
"Feels safe enough to ask in ways that you now feel comfortable"	Human	H	S
"Vulnerability took away the sense of competition: being able to say I'm not good at this is how you can look at getting better"	Claude	S	H

Table 3: Social vs. Human Capitals [10]

5.2 Cultural vs. Social vs. Political Capital

Some classification fluctuation also occurred between Cultural and Social/Political Capital statements. This code confusion echoes the Social vs Human struggle, as Cultural, Social, and Political elements often intertwine in language and require deeper meaning differentiation. Without wider framing and experience in recognizing context cues, Social references can easily be mistaken for Cultural or Political details and vice versa. Unlike more straightforward capital types like Financial or Natural, limited agreement even among human coders for these codes confirms the intricacies of confidently pulling apart Cultural and Social attributes. Nonetheless, being common capital subtypes, better grasping the nuances between them remains important for achieving higher classification competence. The statements where AI models were misclassifying Cultural or Social/Political Capitals are presented below:

Statement	Rater	Rater Code	Agreed Code
"We felt like lab rats when the topic of racism came to the forefront"	Bing	P	C
"Ask questions that you may not have felt comfortable about race and gender etc"	Human	C	S
"Michele and Shelly—literally my whole table said we had one convo with Michele and we were hooked"	Bing	P	S

Table 4: Cultural vs. Social/Political Capitals [10]

5.3 Short/One-Word Statements

Another notable source of disagreement was on short, one-word statements, where the minimal context made consistent classification difficult, even for human coders. With little textual content to analyze, both the Human Annotators and AI systems struggled to reliably categorize these choppy responses into the capital codes. Unlike lengthier statement elaborations with more descriptive cues, single-word phrases lack sufficient semantic and contextual clues to confidently determine an appropriate code. Some of these shorter statements that were coded haphazardly are:

Statement	Human	Claude	Bing
"Gathering"	S	C	S
"Investment"	F	H	F
"Mobility"	X	S	H

Table 5: One-Word Statements [10]

With minimal content to analyze, randomness also plays a role in how the AI models were able to determine sentiment categories – simple chance dictates agreement levels were no supported by true analysis competencies. As such, while inconsistencies persisted across all statement lengths, the severely limited information in one-word examples rendered reliable classification essentially impossible. Further testing focused specifically on performance by input size could help control and quantify this variability. Nonetheless, the struggles on single-word instances underscore fragility applying these coding frameworks on sparse qualitative data inputs.

5.4 Misclassification – Human Annotators with Human vs. Social Capital

Additionally, in several cases the human annotators labeled statements as Human Capital that both AI systems categorized as Social Capital, with high confidence. This exposes potential human bias leaning towards the Human code that could be further investigated. While the original annotators viewed concepts like inspiration, creativity, and relationships as cues denoting Human Capital, the AI models reliably flagged those contexts as Social in nature. With social themes being focused on connections that enable access and exchange, Claude and Bing potentially hold more accurate stances on statements misclassified by the human annotators. Controlled follow-up experiments could determine whether human coders require better guidelines for distinguishing between intellectual, emotional, and personal development facets (Human Capital) rather than externally-focused interactions and impacts (Social Capital). Regardless of final designations, the results highlight likely inconsistent human tendencies when establishing sound benchmarks. The statements that the human annotators coded as Human, but the AI systems agreed were Social are:

Statement	Human	Claude and Bing
"Come as you are"	H	S
"Never felt judged"	H	S
"People who take the time to listen to our story"	H	S

Table 6: Human vs. Social Capitals [10]

5.5 Misclassification – Bing AI with Social vs. Political, Cultural, and Human Capital

The last pattern was that Bing often misclassified clearly Social Capital examples as Political, Cultural, or even Human. This revealed deficiencies in accurately discriminating between societal categories requiring sensitivity to subtle semantic and contextual differences. While the human coders could reliably identify social narratives, Bing struggled to separate those relational elements from other superficially similar concepts. Some of these statements include:

Statement	Human and Claude	Bing
"Crave is come as you are and also take what you want"	S	C
"It was because of crave that I can show up and be honest in this space"	S	P
"In crave lets live and learn from each other and grow"	S	H

Table 7: Bing Misclassifications [10]

Bing codes were often only superficial, lacking in the deeper connections and missing core social signals; it instead interpreted statements through improper lenses. A clear growth area would be to improve categorization competencies between societal categories in order to enhance overall accuracy. Further testing could determine if expanded training on a wider variety of social contexts and topics could nurture the intuition that the Bing AI currently lacks. Regardless, Bing's propensity to misclassify social statements significantly inhibits proper capital classification.

5.6 Scope & Limitations

This research study focuses specifically on comparing the capabilities of two conversational AI systems, Claude and Bing, for a sentiment analysis task. The models were evaluated based on their performance in coding 107 statements from community program transcripts using the Community Capitals framework categories. As such, the scope is limited to a narrow dataset from a single domain. Additionally, only two AI systems (each utilizing two different large language models – Claude and GPT-4) were benchmarked, which does not provide a comprehensive overview of all commercially available models. There are several limitations to this narrow scope that should be acknowledged:

1. Testing was conducted on a small sample of only 107 statements from Crave program interviews. This limited dataset likely does not capture the full diversity of language and sentiment that would be encountered in real-world applications. Larger and more varied datasets could reveal different strengths and weaknesses. Similarly, performance was only gauged within the single domain of community program transcripts. Different contexts, such as customer reviews, social media posts, or workplace communications could require different analysis capabilities.

2. With Claude and Bing being the only models compared, the results do not represent a generalization across other AI systems. The inclusion of models like GPT variants, Llama, Vicuna, or BLOOM could demonstrate different sentiment detection skills. Additional cutting-edge systems that leverage different techniques should be investigated before making broad and official claims. Finally, only the basic out-of-the-box versions of Claude and Bing were tested without any specialized tuning or training. Fine-tuning on textual datasets similar to the evaluation prompts, priming models with topic-specific details, or emotionally appealing to these models could significantly impact analysis aptitudes.

3. Another thing to note is that all of the textual data and AI models used in this study utilized and were trained on American English. Testing on other languages, regional dialects, and cultural norms could reveal different parsing and comprehension capabilities. For instance, models trained on British English jargon may encode cultural nuances distinct from American sentiment patterns. Multilingual model evaluation is required for broad claims about conversational AI abilities.

4. The annotator that was labeled as "expert" was labeled as such because of their expertise in the computer science domain. However, sentiment analysis involves significant subjectivity, so expertise should align with demographics of the textual data source or the necessity for the sentiment analysis. For the community program transcripts, an expert specializing in sociology or community development may possess different judgments. Defining area experts based on dataset origin rather than technical field could enable more appropriate human rating. Another way to go about this would be to have the people who spoke their statements be the ones who encode what sentiment they meant by their statement.

Overall, by focusing on just two models and operating on a narrow dataset, this initial study was limited in scope. Testing across more models on larger, more varied corpuses could strengthen findings by better capturing real-world complexities. Nonetheless, these limitations provide clear guidance for designing expansive follow-on experiments that address key gaps. Broadening data diversity, models evaluated, and customization approaches would build help generalize an understanding of modern conversational AI's sentiment analysis capacities versus human judgement.

6 Conclusion

This study presented a comparative analysis of human and AI performance in coding qualitative data from community development program transcripts. The results demonstrate promising but imperfect agreement between two AI models, Claude and Bing, versus three human annotators and one expert annotator. There are clear gaps in the AI systems' abilities to match human judgment, particularly for statements involving metaphorical language and text that overlaps across multiple categories.

The findings provide a case study for how developers can gauge AI models on sentiment analysis tasks against human baselines. Both Claude and Bing exhibit strengths in matching human codes for certain categories, but also significant areas needing improvement. Testing on larger datasets and providing more surrounding context may improve performance. Most importantly, human-AI comparative studies reveal categories that machines still struggle with compared to humans, highlighting the areas for further research in order for ML systems to reach a similar threshold to human sentiment analysis across all types of language inputs.

Overall, the analysis exposed significant gaps between human and AI classification capabilities on this complex qualitative analysis task:

1. AI models do not agree with themselves; single runs and even aggregate runs are not reliable as they only produce moderate agreement with human annotators and low agreement with the expert annotator. Therefore, we do not have reason to believe that either of the AI models are performing at a comparable level as the expert annotator. The AI models approach the problem just as an undergraduate, inexperienced researcher would approach the problem.
2. Patterns emerged as AI models encountered challenges in processing information when presented with minimal context and metaphorical language. Specific misclassifications include:
    - **Social vs. Human Capital**
    - **Cultural vs. Social vs. Political Capital**
    - **Short or One-Word Statements**
    - **Human Annotators with Human vs. Social Capital**
    - **Bing AI with Social vs. Political, Cultural, and Human Capital**
3. The AI systems demonstrated a reluctance to admit deficiencies by rarely assigning the "None" code compared to the human coders & they were also prone to duplicating labels in consecutive statements, even when these codes were not the logical label; these could be due to poor prompting.

6.1 Future Directions

This research represents an initial exploration into comparing the capabilities of conversational AI systems. As such, there are several promising avenues for further investigation that could shed additional light. One major area warranting future examination is testing Claude and Bing Creative mode on larger, more diverse datasets. The models were only evaluated on a small sample of prompts in this study. Expanding prompts to cover a wider range of topics and statement types would provide more robust quantification of strengths and weaknesses. Additionally, model performance could vary across different domains, so assessing on data from multiple fields is warranted. Along similar lines, evaluating incremental versions of GPT and Claude as they are released over time could provide information about how the codes are affected by the version of the LLM used to conduct analysis. There have already been several major iterations to these models. Comparing early versions to the most cutting edge iterations on the same test prompts could demonstrate tangible results.

Additional models beyond Claude and GPT should be included in future comparative studies. Models such as different versions of GPT, BLOOM, Llama, Vicuna, and other large language models could exhibit different conversational and sentiment analysis capabilities. Expanding the number of models provides a more comprehensive experiment of the commercially available conversational AI ecosystem. Finally, further research should investigate how providing more contextual information could impact analysis performance. For example, supplying examples, emotional appeals [8], or more background details could potentially improve agreement with human ratings under certain prompt types. Too much context could overwhelm models or lead to hallucinated facts. Testing how context effects could provide guidance on optimal information to include. It is important to note that the prompt the AI models were provided consisted of the same 107 statements, the Community Capitals Framework, and instructions on how to code each statement.

In closing, this research sets the stage for multiple avenues of evaluation of commercial conversational AI models. While still providing results, there remains substantial room to build on these initial findings through broader,

more diverse testing; assessing more versions over time; expanding to additional models; and providing better context. The rapid pace of advancement in this domain necessitates continued investigation as new techniques emerge.

7 Acknowledgements

As the author of this paper, I would like to express my gratitude by acknowledging the following individuals and organizations for their support on this research:

- Dr. Daniel Myers, Principal Investigator, for his significant mentorship.
- James McIntyre and James Temple, Co-Researchers, for their collaborative efforts in making this study possible.
- Crave, for sharing their valuable insights through interviews, contributing to the depth of this research.

This research would not have been possible without the support and contributions of the above individuals and organizations.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[2] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300. IEEE, 2019.

[3] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[4] Crave. https://cravefla.org/.

[5] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 75–78, 2023.

[6] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. Llm-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*, 2023.

[7] Hossam Elzayady, Khaled M Badran, and Gouda I Salama. Sentiment analysis on twitter data using apache spark framework. In *2018 13th international conference on computer engineering and systems (ICCES)*, pages 171–176. IEEE, 2018.

[8] Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*, 2023.

[9] M. Haskell, S. Mehdinia, and D. S. Myers. Crave of central florida: A leadership development program for the "spiritually curious". White paper, Rollins College Community Impact Lab, 2021.

[10] Crave and Daniel Myers. Community capitals framework and interview transcript statements, 2021.

[11] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282, 2012.

8 Appendix: Prompt

[10] This is the prompt that was provided to the AI models in order to label each statement:

We are conducting a study on the impacts of a community development program called Crave. The program features small group gatherings, discussions related to goals and spirituality, and professional development workshops.

To assess the impact of Crave, we are using the Community Capitals Framework, which identifies seven dimensions of community resources that play a role in building healthy and sustainable cities and regions. The seven Community Capitals are:

1. **Natural capital.** The natural resources and environment of a particular place, which can include traditional exploitable resources, but also encompasses weather, natural features, beauty, and geographic location.

2. **Cultural capital.** The traditions, language, heritage, and other expressions of the peoples that live in a community.
3. **Human capital.** The ability for community members to acquire new resources, develop knowledge, and pursue opportunities. This capital also encompasses the values of inclusive and proactive community leadership.
4. **Social capital.** Connections between individuals and organizations that contribute to development. Bonding social capitals are tighter connections that build cohesion within a community. Bridging social capitals are looser ties that connect different organizations and subgroups.
5. **Political capital.** Access to centers of power, government officials, and the political process. This capital also encompasses the capacity of community members to advocate for their own interests.
6. **Financial capital.** Assets available to support capacity-building, including investments in business, social entrepreneurship, and future community development.
7. **Built capital.** Physical infrastructure that supports the other capitals, including buildings, transportation, utilities, and digital connectivity.
8. **None (X).** You may choose "None" if you feel that the item does not clearly relate to any of the Capitals.

The bullet points below are a series of quotes and observations taken from a group interview with Crave participants. For each item, choose the most appropriate Community Capital. Choose one and only one Capital for each item.

Crave Group Discussion Bulleted Notes: Certainly! Here is a complete LaTeX document that you can use for your appendix:

1. "I forgot until now the fact that we gathered in one place together"
2. "Gathering"
3. "Time at the little red house—small space with informal meetings, super spiritual and vulnerable"
4. "everybody in crave showed up with true selves"
5. "The power and influence of Michele– Only took one conversation with her and we bought into it"
6. "Vulnerability—being in a safe place, being authentic"
7. "Little red house talks, quickly where people got to all speak about what we believe in spiritually"
8. "The fact that we felt comfortable enough to share that is really meaningful"
9. "Vulnerability was meaningful there and led me to being super vulnerable throughout"
10. "Black/African-American experiences—culture shocks, able to address these but not always get the answers"
11. "Culture shocks: gather with food, food that everyone eats is not the same, asking the invitees what you would like"
12. "What's up with the Rabbit Food??"
13. "Speakers with more privilege and had more access to resources"
14. "The way they solved problems made us feel angered that we couldn't do that"
15. "We felt like lab rats when the topic of racism came to the forefront"
16. "The way we were raised as Christian—could not associate drinking in the church"
17. "Now with Crave we can do things like that "Pints and parables"
18. "Beer and wine before dinner"
19. "I always felt heard, I always felt respected"
20. "For the first time, "woah I hear what you're saying and even though I don't have the answer I value that"
21. "Come as you are"
22. "Sharing spirituality and feeling where you are in life and what I might believe and how that connects with that"
23. "Safety in this space: previous experience vs now"
24. "In Crave, let's live and learn from each other and grow"
25. "We don't shy away from the discomfort"

26. "To be welcome in a Sunday school class of White people that wanted to hear what you had to say"
27. "Situations that open us up to relationships"
28. "People who take the time to listen to our story"
29. "Ask questions that you may not have felt comfortable about race and gender etc"
30. "Feels safe enough to ask in ways that you now feel comfortable"
31. "Never felt judged"
32. "No one is out here to hurt anyone, just trying to learn and get better and do good"
33. "Understanding of whether our belief system or culture is different, we're still on the same journey"
34. "Let go of the outcomes or specifics and just know that there is a journey"
35. "How does the idea of being in a safe group translate into supporting your professional life?"
36. "Vulnerability took away the sense of competition: being able to say I'm not good at this is how you can look at getting better"
37. "You're still developing"
38. "Gets rid of all the bullsh*t"
39. "You don't have to present or code switch; you get to the real and get better at the real"
40. "When we know how to meet people's needs then we can bridge the gap, when you see and share the vulnerability you can help each other grow"
41. "Other Crave leaders have helped me so much"
42. "We have all built Crave together over the years"
43. "Help from peer to peer vs only from the expert giving the information"
44. "Connection with our group that really felt like it was something different"
45. "I loved my group… when it came to my leaders we became like sisters and brothers"
46. "Felt like family"
47. "Helping us to gain the confidence to be the experts that we are"
48. "It wasn't until I went through Crave until I was able to confidently say "yes I am a leader"
49. "confidence comes from the competence"
50. "That's what makes the growth equitable, throughout Crave I felt like we created the path… that let me know that as the leaders we can create the pathway in terms of where we go"
51. "When we left we could get what we needed, everyone made sure that they got what they needed from the situation"
52. "Even the facilitator grew through the process"
53. "we created a path"
54. "The group is always figuring out what the group needs"
55. "Tracking data and measurements and finances"
56. "It's easy to start with your heart, leader encouraged people to have the viable and sustainable pieces"
57. "Allison Fairfax was about telling your story, using your story elements into something structured so you have a purpose about what you shared"
58. "Telling your story in a way that people care—that was so impactful"
59. "Prof dev sessions helped me to think about things differently—find the difference between strategy and operation"
60. "Define the problem and brainstorming vs action"
61. "Try to understand how different people think"
62. "Prof dev science on anger—when I heard his story and the things that he went through, the speaker had anger and transformed it"
63. "Taking energy and transforming it into something that you want—channel anger to get the outcome that you want"

64. "MLK on the outside, Malcolm X on the inside" get the outcome you want"
65. "Coming to Crave"
66. "I was in a place of timid/don't know how to step into it"
67. "I'm in a bad place and want to turn this around"
68. "Fear of not knowing how to make ends meet"
69. "People who want to support you is the difference between sink and swim"
70. "Linked to access? Connected to fear and a different thing"
71. "Thinking about belonging? What "bucket"
72. "Crave gave me hope, hope is a remedy to fear"
73. "The emotions were accepted and we could talk about them, what is the emotion here and what am I struggling with, that is not really a lot of communities that allow us to explore them"
74. "Church for me has been one of shame, there is no growth in that, it's all direction"
75. "In traditional professional development place it's all logical, there is no space for creative and vulnerable processes"
76. "It's such a fixed mindset when you think about church and tradition, "church feels fake to me, it's built off of rules but rules then change" think like LGBTQ, drinking, being black—all these things that were once wrong are now right? And it's built off of what's best for people's pockets, not what's best for people."
77. "That's one thing that separates crave from other ministry possibilities, get to see real work—work that's not doing harm and not perpetuating harm"
78. "Crave is come as you are and also take what you want"
79. "We've all had different professional experiences and have taken different things from each"
80. "Take it or leave it"
81. "Feeling safe enough to say thanks for bringing people in"
82. "Being listened to and heard—saying things like "we could have benefited from this more" and seeing things change in real time"
83. "If the organization is willing and able to listen and actually change, then I feel like I belong because I'm respected and I have access to this organization—allow us to make space for people with other gender identities"
84. "I think about power, how many times that we were able to push back against white male middle class construct, when I'm with crave I feel like people are going to be able to hear me"
85. "I don't know if I could've done peace and justice institute without crave"
86. "Diversity committee (old white people) "I have never been so honest with white people in my life"
87. "It was because of crave that I can show up and be honest in this space"
88. "Then we think about Now What—shows Crave's willingness to continue to grow, we don't have all the answers, always being challenged to grow"
89. "Mobility"
90. "Motivated us to make a change"
91. "Michele and Shelly—literally my whole table said we had one convo with Michele and we were hooked"
92. "We talked about our lives, our passions, she didn't talk about herself or crave"
93. "Crave is about YOU and YOUR project, you create the collaborative growth etc"
94. "God is doing this through me"
95. "I've seen time and time again how this network has done the big and the small"
96. "I didn't believe in that before Crave, I didn't believe that people would just help you for no reason"
97. "I didn't believe that people genuinely cared without getting something out of it"
98. "To say charity is a part of it does not fit at all, it's so much about individuality"
99. "Investment"

100. "Platform for promotion—you're allowed to be a beginner, you're allowed to not know and you're allowed to not have it all figured out, willingness to be a beginner and not know all the answers"
101. "People are longing for the collaboration—now that crave has existed for 4 years, so many people are talking about it, are being excited about it—makes me believe God is still working to lead this group together"
102. "You don't necessarily know what you're going to get out of things but in this space you're allowed to grow, that's what differs us from other professional development spaces"
103. "Life changing Connections made from the network, from the presenters"
104. "ELAR institute, mission increase, victory cup,"
105. "Internally we assisted each other with our projects and stuff"
106. "Self-care aspect, how to tend to your own needs and connect to yourself so that you can show up and keep showing up, how you can keep doing that work"
107. "Love! We have to love what you do and yourself to be able to serve others"
