

SCALE: Towards Collaborative Content Analysis in Social Science with Large Language Model Agents and Human Intervention

Chengshuai Zhao[♣], Zhen Tan[♣], Chau-Wai Wong[♥], Xinyan Zhao[♣], Tianlong Chen[♣], Huan Liu[♣]

[♣]Arizona State University, [♥]NC State University,

[♣]University of North Carolina at Chapel Hill

{czhao93, ztan36, huanliu}@asu.edu, chauwai.wong@ncsu.edu,
ezhao@unc.edu, tianlong@cs.unc.edu

Abstract

Content analysis breaks down complex and unstructured texts into theory-informed numerical categories. Particularly, in social science, this process usually relies on multiple rounds of manual annotation, domain expert discussion, and rule-based refinement. In this paper, we introduce SCALE,¹ a novel multi-agent framework that effectively Simulates Content Analysis via Large language model ag^Ents. SCALE imitates key phases of content analysis, including text coding,² collaborative discussion, and dynamic codebook evolution, capturing the reflective depth and adaptive discussions of human researchers. Furthermore, by integrating diverse modes of human intervention, SCALE is augmented with expert input to further enhance its performance. Extensive evaluations on real-world datasets demonstrate that SCALE achieves human-approximated performance across various complex content analysis tasks, offering an innovative potential for future social science research.

1 Introduction

Content analysis is a key research method in various disciplines (Benoit, 2014; Dart, 2014; Macnamara, 2005; Hara et al., 2000). It breaks down complex and unstructured text into numeric categories based on theory-driven rules (Krippendorff, 2018; Weber, 1990), offering a systematic and quantitative approach to interpreting sophisticated information (Holsti, 1969; Neuendorf, 2017; Riffe et al., 2023). Particularly, in social science, content analysis is usually labor-intensive and time-consuming (Hopkins and King, 2010; Zhao and Wong, 2024). It often requires a team of researchers to manually annotate sizable datasets, conduct comprehensive discussion, and iteratively refine coding rules (i.e., codebook) in multiple rounds to ensure

¹GitHub: <https://github.com/ChengshuaiZhao0/SCALE>

²Coding refers to the process of converting text entries into categories, similar to text classification tasks.

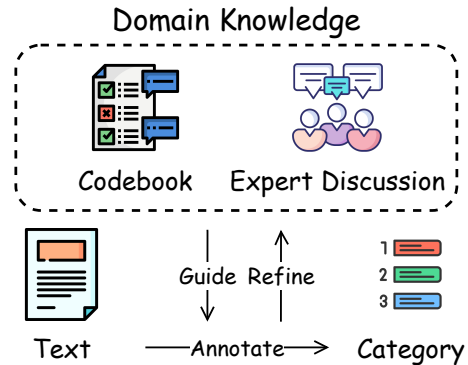


Figure 1: Illustration of content analysis. Social scientists convert text into categories based on a codebook, discuss discrepancies, and refine the codebook.

reliability and validity of findings (Cohen, 1960; Krippendorff, 2018; Riffe et al., 2023), as illustrated in Figure 1. This artificial process, while rigorous, presents two challenges. First, it relies heavily on domain-specific knowledge and individual scientists, potentially introducing subjectivity and limiting generalizability. Second, the substantial human resources demanded by content analysis tasks make it difficult to scale, especially as the volume of digital data expands exponentially.

Recent years have witnessed significant progress in artificial intelligence (AI), especially with the advent of large language models (LLMs) (Ziems et al., 2024; Zhao et al., 2024; Xia et al., 2023; Zhao et al., 2023; Jiang et al., 2024). LLM agents emerge as a versatile tool across a wide range of domain-specific tasks (Guo et al., 2024; Wang et al., 2024; Xi et al., 2025; Park et al., 2023; Li et al., 2024b; Wu et al., 2024). However, content analysis tasks present unique challenges, requiring a fine-grained understanding of social science principles, human-like collaborative interaction, and rule-based iterative refinement, which limits the effectiveness of agent systems derived from other tasks.

In this paper, we propose a novel multi-agent

framework to **Simulate Content Analysis** via **LLM agENTS** (SCALE) detailed in Figure 2. SCALE imitates key phases of content analysis, including text coding, collaborative discussions, and dynamic codebook evolution, while capturing the reflective depth and adaptive discussions of diverse human researchers, thereby reducing subjectivity and improving scalability. Moreover, by incorporating different human-AI collaboration modes inspired by social influence theories (Cialdini and Cialdini, 2007; French, 1959) and human-computer interaction theories (Suchman, 1987; Sundar, 2020), our framework extends multi-agent interactions with human expert intervention. This potentially mitigates algorithmic bias and strengthens contextual sensitivity, making it suitable for a wide range of social science content analysis tasks.

Comprehensive experiments on multiple real-world datasets demonstrate that SCALE is versatile across diverse social science contexts and has the ability to approximate human judgment in complex content analysis tasks. Developed in collaboration with social scientists, we demonstrate the potential of our framework to transform content analysis in the social sciences and humanities, inspiring the integration of AI into other domain-expert-dependent research methodologies. We summarize our contributions as follows.

- ★ **Scalability Enabler.** By harnessing the generative power of LLMs, our proposed SCALE significantly reduces the time, human resources, and costs traditionally required for content analysis tasks, enabling large-scale and high-quality annotation. To the best of our knowledge, this is the first work to capture and simulate the rigorous and dynamic process of quantitative content analysis for social science via LLMs.
- ★ **Praxis-Informed Design.** SCALE’s design incorporates the domain knowledge of social science through the deep involvement of social scientists. Its key phases—independent text coding, collaborative discussions, and dynamic codebook evolution—faithfully reflect the principles and standards of manual content analysis while being implemented within a computing framework.
- ★ **Human Intervention.** Our framework provides a portal for domain experts to intervene with custom scopes (i.e., *targeted* or *extensive* intervention) and roles (i.e., *collaborative* or *directive* intervention). Incorporating expert input aug-

ments AI decision-making by mitigating LLM bias and aligning with expert oversight.

- ★ **Extensive Validation.** SCALE demonstrates effectiveness across content analysis tasks involving diverse topics (e.g., analysis of emotion, perspective, and dialogue). Our comprehensive experimental evaluations and analyses by domain experts confirm that SCALE can closely mimic human judgment in content analysis, delivering automated, valid, and reliable results invaluable for large-scale social science research.

2 Related Work

Content Analysis. Content analysis has long been a foundational method in the social sciences and humanities, providing a structured approach to converting qualitative text into quantitative data (Krippendorff, 2018; Neuendorf, 2017; Riffe et al., 2023). Recently, content analysis has significantly advanced the understanding of complex social issues, ranging from political polarization (Conover et al., 2011) to emotional contagion (Kramer et al., 2014) and group dynamics (Holsti, 1969). These traditional methods rely on manual annotation by human coders, who use predefined rules in the codebook to categorize text, often iteratively refining their coding schemes in multiple rounds of discussions (Riffe et al., 2023). Although manual content analysis provides robust and theory-driven insights, it remains labor-intensive, time-consuming, and prone to subjectivity (Hopkins and King, 2010). Furthermore, as the volume of digital text increases, scaling traditional methods to accommodate larger datasets has become increasingly challenging (Zhao and Wong, 2024). The advent of powerful AI and LLM offers an automated and more scalable solution (Eloundou et al., 2023; Achiam et al., 2023; Li et al., 2024a).

Multi-agent Systems for Social Science. Multi-agent systems (MAS) have become increasingly prevalent in computational social science, modeling social phenomena through agent individuals or groups with predefined behaviors or decision-making rules (Van der Hoek and Wooldridge, 2008; Chen et al., 2021; Chmura and Pitz, 2007; Macal, 2016; Lee et al., 2018; Chen et al., 2018; Dehkordi et al., 2023; Uhrmacher and Weyns, 2018). Recent MAS explore to simulate human-like deliberation for more nuanced decision-making such as data interpretation (Gürçan, 2024; Turgut and Bozdog, 2023). However, existing systems often lack dy-

dynamic evolution of decision rules, collaborative interactions, and a systematic comprehension of social science principles (Gheyle and Jacobs, 2017), which cannot reflect the sophisticated and theory-informed process of content analysis.

Human Intervention. Human intervention remains essential for the reliable deployment of AI-driven systems (Renner, 2020; Shoshitaishvili et al., 2017). As a general framework, human-in-the-loop (HITL) systems allow experts to refine AI outputs, ensuring alignment with domain-specific knowledge and mitigating algorithmic bias (Mosqueira-Rey et al., 2023; Ghai and Mueller, 2022; Xu et al., 2023; Jolfaei et al., 2022; Wu et al., 2022; Zanzotto, 2019). This is particularly important in social sciences and humanities, where interpretative depth and contextual sensitivity are critical (Dautenhahn, 1998; Goodsell, 2013). Recent approaches (Arambepola and Munasinghe, 2021) integrate expert feedback to adjust categories or schemes iteratively. Our framework significantly extends this line of work by designing different modes of human-AI collaboration informed by social influence theories (Cialdini and Cialdini, 2007; French, 1959) and human-computer interaction theories (Suchman, 1987; Sundar, 2020).

3 Traditional Content Analysis in Social Science: A Preliminary

Social scientists/coders conduct content analysis by manually annotating textual data to uncover potential patterns and insights. A group of $K \geq 2$ coders indexed by $k \in \{1, \dots, K\}$ first develop an initial codebook $\mathcal{C}^{(0)}$ that contains a set of coding rules grounded in relevant social science theories and contextualized within the given text corpus $\mathcal{D}$. Guided by the codebook, the k th coder independently labels a small set $\mathcal{D}_{\text{tr}} \subset \mathcal{D}$ of text entries (e.g., $n_{\text{tr}} = 10$ to 20) to obtain all coded labels $\mathcal{Y}_k^{(0)} = E_k(\mathcal{D}_{\text{tr}}; \mathcal{C}^{(0)}) \in \{0, 1, \dots\}^{n_{\text{tr}}}$. Later, all coders meet to discuss and resolve inconsistencies, namely, $[\{\tilde{\mathcal{Y}}_k^{(0)}\}_{k=1}^K, \mathcal{C}^{(1)}] = \text{Discuss}(\{\mathcal{Y}_k^{(0)}\}_{k=1}^K; \mathcal{D}_{\text{tr}}, \mathcal{C}^{(0)})$, where $\tilde{\mathcal{Y}}_k^{(0)}$ contains the updated labels by the k th coder and $\mathcal{C}^{(1)}$ is the improved codebook with more specific coding rules and examples for subsequent annotation. This process iterates for multiple (e.g., 3–5) rounds until labels from all coders converge. Finally, the newest codebook is applied to code the whole corpus. Content analysis in this paper centers on two objectives: (i) precisely annotating all text entries and (ii) craft-

ing a well-honed codebook with distinct coding rules.

4 The Proposed Framework: SCALE

4.1 Content Analysis Simulation

We introduce SCALE as a framework that mirrors the key phases of real-world content analysis—text coding, collaborative discussions, and dynamic codebook evolution. The method unfolds in four primary steps summarized in Figure 2. (Figure 5 of Appendix A shows a more detailed version.)

Coder Simulation. Prior to the content analysis task, we set up both LLM agents and an initial codebook. As illustrated in Figure 2(a), we begin by configuring N LLM agents $\mathcal{A} = \{a_i\}_{i=1}^N$, each emulating a seasoned social scientist through a system prompt that incorporates N distinct personas, $\mathcal{P} = \{p_i\}_{i=1}^N$. These personas—based on real-world social scientists—facilitate authentic role-playing. Depending on the specific content analysis task (detailed in Section 5.1), we initialize a corresponding codebook $\mathcal{C}$ that either starts with N' human-expert predefined rules (shown in red) $\mathcal{C} = \{r_i\}_{i=1}^{N'}$ or as an empty set $\emptyset$ prompting agents to propose and iteratively refine the codebook from scratch. For simplicity, each rule is tailored to cover a single scenario, enabling the categorization of text into a discrete class. The personas for diverse social scientists are provided in Appendix D.1.

Bot Annotation. LLM agents convert text entries into numerical categories by applying theory-informed rules from the codebook. Figure 2(b) illustrates this process. Each agent is assigned an identical batch of B text entries (highlighted in blue) from the text dataset and works autonomously to classify each entry into a discrete category (marked in purple). Mimicking the independent coding approach of human researchers, these LLM agents adhere strictly to the codebook guidelines, which is facilitated by a prompt (see Appendix D.2).

Agent Discussion. In this phase, agents engage in collaborative discussions to resolve discrepancies in their coding outputs as illustrated in Figure 2(c). Due to the initial ambiguity of the evolving codebook and the distinct personas embodied by each agent, it is not uncommon for agents to generate differing annotations for the same text, which mirrors the subjective nature of real-world content analysis. Whenever an agent’s coding diverges from the

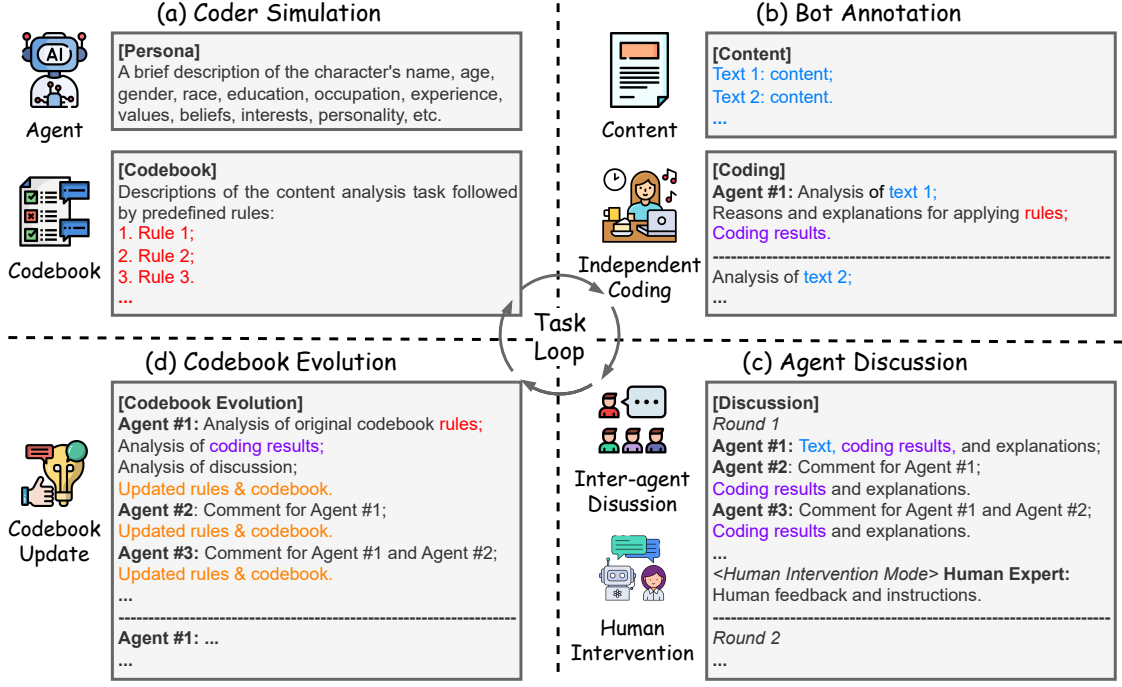


Figure 2: Proposed SCALE framework. (a) Coder Simulation. Initialize LLM agents with real-world personas and set up the codebook using predefined rules. (b) Bot Annotation. Each LLM agent independently annotates a batch of text entries into theory-informed categories based on the codebook. (c) Agent Discussion. Agents participate in multi-round discussion to resolve any inconsistencies in their annotations. Optional human intervention can be integrated to provide feedback and additional instructions. (d) Codebook Evolution. Leveraging the outcomes from the annotation and discussion phases, agents collaboratively and iteratively update the codebook. The refined codebook is then employed in subsequent iterations of the workflow. Data flows are tracked using color-coded text: **Red text**: codebook and rules; **Blue text**: text entries; **Purple text**: coding results; **Orange text**: codebook evolution. A comprehensive illustration of SCALE framework can be found in Figure 5 and Appendix A.

consensus, the agents initiate a structured discussion up to a maximum of K rounds. During each round, they update their annotations along with explanations based on peer opinions until they converge on a unanimous decision or reach the limit on discussion rounds. Once a text entry is finalized, the agents will move to the next entry. The prompt used for the discussion phase is provided in Appendix D.3.

Codebook Evolution. In this stage, agents refine the codebook by incorporating insights from their discussion as illustrated in Figure 2(d). As noted earlier, the initial rules can be ambiguous, overlapping, or insufficiently detailed to cover all categories. To address these issues, we introduce an iterative codebook evolution process grounded in domain expertise. Specifically, the evolution approach offers two strategies: one enriches existing rules by adding clarifying examples and explanations, whereas the other allows for adding, removing, or modifying rules to dynamically adjust the set of categories. In practice, agents first propose an updated codebook, then engage in multiple rounds

of discussion to refine it until consensus is reached (indicated in orange). The finalized codebook subsequently guides the next task cycle. Importantly, agents may also retain the current codebook if no changes are warranted. The delicate prompt for the codebook evolution is provided in Appendix D.4.

4.2 Human Intervention

We further augment SCALE by integrating diverse human intervention modes [see Figure 2(c)] that empower domain experts to provide targeted feedback and instructions. Depending on the scope that human experts can control, human intervention can be categorized as *targeted* versus *extensive*. Further, intervention can be formulated as *collaborative* or *directive*, based on the role the human experts play. Intervention modes are detailed below.

- **Targeted Intervention.** The scope of intervention is limited to the agent discussion phase. The process is done with less human oversight.
- **Extensive Intervention.** Human intervention can be applied to both the discussion and code-

Table 1: List of used datasets and evaluated content analysis tasks.

Dataset	Variable(s) to be Coded	# Texts	Coding Type	# Classes
Brand Consumer Dialogue (BCD)	Primary Topic (PT)	92	Multi-Class*	10
	Dialogue (D)		Multi-Label†	7
Cancer Emotional Support (CES)	Emotional Support (ES)	40	Multi-Class	3
Cancer Narratives (CN)	Narrative Event Sequence (NES)	60	Multi-Label	5
	Narrator Perspective (NP)		Multi-Class	5
Flint Water Poisoning Emotion (FWPE)	Emotion (E)	100	Multi-Label	13
Product Incidents Sentiment (PIS)	Sentiment (S)	200	Multi-Class	3

* Multi-Class: each sample is assigned exactly one label from a set of mutually exclusive categories;

† Multi-Label: each sample can be assigned multiple labels from a set of non-exclusive categories.

book update phases. While it may slow automation and increase costs, it ensures that AI discussions remain closely aligned with expert insights.

- **Collaborative Intervention.** Human experts are involved as collaborators. LLM agents may either accept or reject feedback and suggestions from human experts, which fosters an interactive and cooperative discussion loop.
- **Directive Intervention.** Under this mode, human experts serve as absolute authority. LLM agents must adhere to every instruction provided, thereby establishing a highly prescriptive and unequivocally top-down approach.

It is noted that, in practice, we can combine various scopes and roles of expert control to enable custom interventions (e.g., *targeted-collaborative* and *extensive-directive* intervention). Different human interventions are implemented by delicately crafted prompts provided in Appendix D.5.

5 Experiments

5.1 Datasets and Tasks

Our experiments leverage five real-world datasets, each meticulously annotated and validated by social science experts. These datasets are organized into seven distinct tasks that encompass both multi-class and multi-label classification challenges. Table 1 summarizes their key characteristics, with further details provided in Appendix B.1.

5.2 Experiment Settings & Metrics

Experiment Settings. Our multi-agent system is built on GPT-4o and GPT-4o-mini. While alternative backbones (e.g., Gemini, Claude) are available, our focus is on assessing whether LLM agents can simulate sophisticated social science tasks rather than comparing various LLMs. For each backbone, we explore four prompting

strategies: vanilla, chain-of-thought (COT) (Wei et al., 2022), tree-of-thought (TOT) (Yao et al., 2024), and self-consistency (Wang et al., 2023). The prompts for CoT and ToT are provided in Appendix D.6. The identifiers for GPT-4o and GPT-4o-mini are gpt-4o-2024-05-13 and gpt-4o-mini-2024-07-18, respectively. We simulate a real-world scenario in content analysis by setting $N = 2$ agents, a batch of $B = 20$ texts, and $K = 3$ discussion rounds. We consider and discuss more hyperparameters in Section 5.4.1.

Metrics. We define the following evaluation metrics for our content analysis tasks. For various classification tasks, we employ standard multi-class classification accuracy and define the multi-label classification accuracy as $\text{ACC} = 1 - \text{Hamming Loss}$. Furthermore, we use the agreement rate—defined as the proportion of text entries where all agents concur—to assess the level of consensus during discussions. All experiments were conducted over 10 independent repetitions, with the average results reported to ensure robustness. Further metric details are provided in Appendix B.2.

5.3 Superior Performance of SCALE

5.3.1 Automatic Content Analysis

We first assess the performance of automatic content analysis without human intervention. Table 2 reveals that SCALE using GPT-4o achieves satisfactory results across a diverse range of tasks and prompting techniques, with an average accuracy of 0.71. Notably, different prompting techniques offer distinct benefits: compared to the vanilla model, self-consistency prompts boost labeling accuracy by 3.2%. However, in certain cases (e.g., CN-NP), COT prompts lead to a significant performance drop. We attribute this decline to the challenging nature of tasks with ambiguous coding rules that introduce greater subjectivity, where the step-by-

Table 2: Coding accuracy across various tasks and backbones using automatic framework.

Backbone (w/o intervention)	BCD-PT	BCD-D	CES	CN-NES	CN-NP	FWPE	PIS	Average
GPT-4o	0.51	0.54	0.63	0.79	0.77	0.91	0.88	0.72
GPT-4o w/ COT	0.54	0.61	0.58	0.71	0.52	0.90	0.85	0.67
GPT-4o w/ TOT	0.57*	0.63	0.53	0.73	0.70	0.88	0.87	0.70
GPT-4o w/ self-consistency	0.51	0.57	0.65	0.80	0.83	0.92	0.91	0.74
GPT-4o-mini	0.38	0.47	0.58	0.73	0.55	0.79	0.82	0.62
GPT-4o-mini w/ COT	0.19	0.47	0.53	0.72	0.43	0.81	0.71	0.55
GPT-4o-mini w/ TOT	0.35	0.48	0.58	0.83	0.70	0.84	0.84	0.66
GPT-4o-mini w/ self-consistency	0.43	0.50	0.58	0.79	0.72	0.85	0.87	0.68

* Bold values indicate the best performance of all four backbones.

Table 3: Coding accuracy across various tasks using the human-intervention framework.

Intervention Mode		CES	CN-NES	CN-NP	FWPE
No Intervention*		0.63	0.79	0.77	0.91
Targeted	Collabo.	0.73	0.89	0.87	0.95
	Directive	0.73	0.85	0.87	0.95
Extensive	Collabo.	0.77	0.89	0.90	0.96
	Directive	0.77	0.91	0.97	0.96

* Automatic mode, same as in the first row of Table 2.

step thinking induced by COT may inadvertently undermine performance. Furthermore, as expected, GPT-4o outperforms its distilled version, GPT-4o-mini, by an average margin of 13.6%. Additionally, we evaluate the impact of inter-agent discussion on coding accuracy. As detailed in Table 4 and Appendix C.1, omitting the collaborative discussion phase results in a 14.2% reduction in terms of average annotation accuracy. A comprehensive analysis of collaborative inter-agent discussion is elaborated in Section 5.4.2.

5.3.2 Human-Intervened Content Analysis

SCALE can be augmented with the four intervention modes listed in Table 3. The automatic content analysis results with human intervention achieve an average accuracy of 0.872, showing an average improvement of 12.6% over content analysis without human intervention, demonstrating the importance of human intervention. Specifically, interventions in the *directive* mode prove more effective than those in the *collaborative* mode, leading to a 13.1% increase in coding accuracy. Similarly, *extensive* interventions generally outperform *targeted* interventions, yielding a 15% average improvement. This aligns with intuition, as human experts can contribute more domain knowledge through intervention across a broader range of agent behaviors in a mandatory manner. Lastly, the benefits of human intervention also vary by task. For example, the

CES task sees a 19% improvement, whereas the FWPE task shows a 4.9% relative gain—likely because the CES task is more domain-specific and subjective, thus more responsive to expert insights. Additional results of SCALE framework before the agent discussion phase are provided in Table 5 and Appendix C.2.

5.4 Extra Investigations and Case Studies³

5.4.1 Q1: What designs can enhance content analysis performance of SCALE?

To answer Q1, we analyze how SCALE enhances content analysis tasks by considering the number of texts, discussion rounds, and agents.

Number of Texts. We first evaluate the influence of the number of texts, B . Figure 3(a) reveals that a moderate B (e.g., 10 and 20) produces the best accuracy. When B is small (e.g., 1), agents frequently refine the codebook, resulting in unstable coding outcomes. However, when B is large, results become more stable, but the overall performance decreases due to less frequent codebook evolution.

Number of Discussion Rounds. Next, we examine the effect of the number of discussion rounds, K . Figure 3(b) reveals that SCALE achieves better performance with $K \geq 3$ as more rounds better facilitate consensus among agents, thereby improving coding accuracy. Importantly, setting K to 0 (i.e., no discussion phase) results in a significant drop in accuracy for several tasks (e.g., BCD-D and CN-NP in Appendix C.3), highlighting the importance of inter-agent discussion.

Number of Agents. Finally, we assess the impact of the number of agents, N . Figure 3(c) reveals that in general, increasing the number of agents improves coding accuracy, as more agents bring di-

³Automatic framework is used for extra experiments.

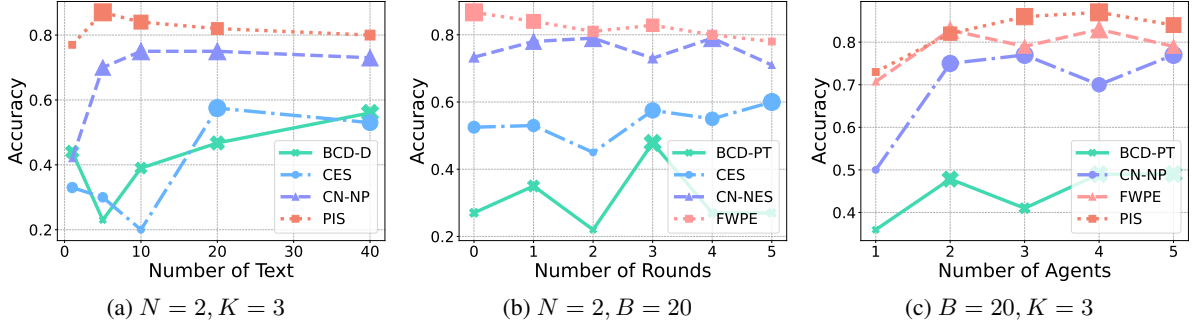


Figure 3: Parameter sensitivity. SCALE shows versatility under different parameter settings.

verse perspectives, fostering more comprehensive discussions. When $N = 1$, SCALE degenerates to a single-agent system, where a single agent performs the coding task without collaboration. As expected, this setup yields the worst performance, verifying the effectiveness of multi-agent design in the proposed SCALE framework.

5.4.2 Q2: How does the discussion between LLM agents impact coding results?

To answer Q2, we conducted a discussion analysis using both qualitative statistics and illustrative examples. Figure 4(a) reveals that inter-agent discussions substantially boost consensus—enhancing the average agreement rate by 41.1% across all seven tasks and overall content analysis accuracy by 15.4%. Similar trends were observed for GPT-4o-mini agents, as illustrated in Figure 7 and Appendix C.4.

An example of sentimental analysis (multi-class classification task) can be found in the PIS dataset: a tweet such as “Hey @SamsungMobileUS, bf has a recalled #GalaxyNote7. Can’t find a replacement S7 Edge in Orlando, FL area. Any ideas or help please?” initially resulted in conflicting sentiment annotations (neutral versus negative). After three rounds of collaborative discussion, both LLM agents agreed on a neutral sentiment—aligning with the ground truth. The complete example is showcased in Appendix E.1.

However, the benefits of discussion can be marginal when agents remain firmly entrenched in their views. In some datasets (e.g., FWPE and PIS), the increase in agreement was less than 3%. For example, in one CES (multi-class classification) task, a Facebook comment “This is so sad :(she was beautiful inside and out! Loved watching her perform <3” sparked a debate: one agent rated the emotional support as moderate (i.e., category 2), while the other deemed it high (i.e., category 3),

and despite extended discussion, their judgments did not converge, as shown Appendix E.2.

The aforementioned instances show that even with discussion, task performance gain can be limited when agents are entrenched in their stances, which might be an innate characteristic of LLMs and influenced by the customized agent persona and background. A moderate level of agent difference in terms of personas, compared to low or high levels, may be most productive in revealing diverse viewpoints and fostering discussion that more likely leads to the “truth” through meaningful exchanges instead of impasse or blind agreement.

5.4.3 Q3: How reliable are the codebooks proposed by LLM social scientists?

One of the key objectives of our method is to refine the codebook for annotating large-scale text. To answer Q3, we analyze the codebook evolution in SCALE. We observe that LLM agents can enhance codebooks in non-structural ways by adding clarifying details and examples. For example, during the PIS codebook update (shown in Appendix E.3), Agent #1 proposed incorporating examples for each sentiment category (positive, neutral, negative) to ensure consistent interpretation, while Agent #2 initially preferred the original version. After discussion, the final codebook merged Agent #1’s detailed examples with Agent #2’s simplicity, achieving a balance of clarity and reliability. This iterative process mirrors core content analysis practices by fostering convergence in agent judgments.

However, the agents were less adept at adjusting codebook categories. For example, in all rounds of FWPE (multi-label classification task) codebook evolutions, both agents maintained that the categorization of twelve discrete emotions (e.g., anger, sadness, and hope) was appropriate, diverging from human experts who ultimately dropped two categories due to overlapping semantic boundaries.

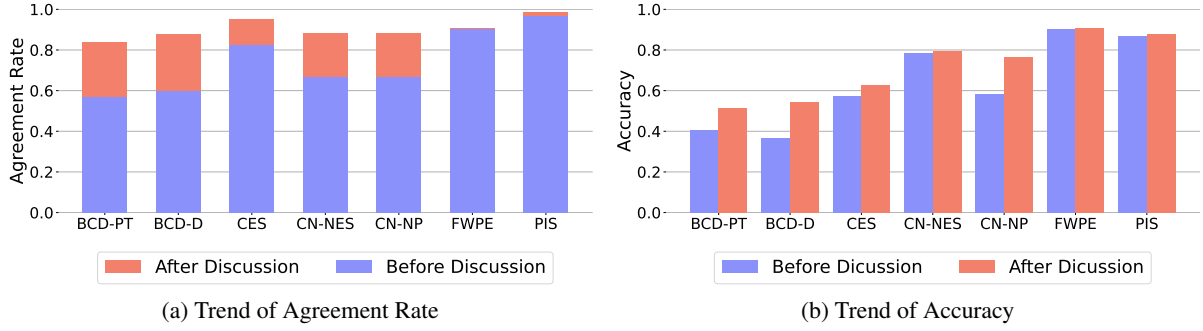


Figure 4: Discussion Analysis. (a) Agreement rate and (b) accuracy are evaluated before and after the agent discussion phase. The inter-agent discussion can promote consensus and further improve coding accuracy.

This limitation might stem from the reliance on predefined rules and patterns in their training data. LLMs may lack additional domain knowledge apart from content analysis tasks to detect subtle conceptual overlaps (e.g., between anger and disappointment or between happiness and pride), leading to rigid adherence to existing category structures and conceptual boundaries. Human experts, on the other hand, can apply more domain knowledge (e.g., the appraisal theory) and theory-based, contextualized reasoning to recognize subtle distinctions between categories, identify overlap, and even add or drop new categories when necessary.

5.4.4 Q4: To what extent can LLM agents simulate content analysis?

To answer Q4, we examine the complete workflow of SCALE on the NES (multi-label classification) task. Here, two agents mimic social scientists to classify cancer narrative events into multiple categories (e.g., prevention, detection, treatment, survivorship), which is reported in Appendix E.4.

Text Coding. Both agents independently applied the codebook rules to annotate the presence of one or more cancer narratives. For example, when annotating the text “When I hear that some women feel too afraid to go for a mammogram...,” both agents agreed on labeling it as detection. In contrast, for the text “...After that I will have 25 days of radiation... But through it all, I have had great support from my family and friends,” Agent #1 focused solely on treatment, while Agent #2 identified both treatment and survivorship. Overall, the initial coding yielded a 66.7% agreement rate.

Collaborative Discussion. Following independent coding, the agents discussed their differing annotations. In the example, after three rounds of collaborative discussion, they converged on two nar-

rative events—treatment (chemotherapy, radiation) and survivorship (support from family and friends). Across tasks, such discussions resolved 21.7% of initial disagreements, underscoring the value of shared interpretation.

Dynamic Codebook Evolution. After each discussion round, agents reviewed and refined the codebook to enhance clarity. For instance, in the first round they enriched the “survivorship” category by adding examples that distinguished post-treatment narratives from ongoing medical interventions. These refinements, which aligned with human expert updates, reduced ambiguity and improved subsequent coding accuracy.

This case study generally demonstrates that our algorithm not only mirrors the iterative and theory-driven process of traditional content analysis but also produces human-approximated performance.

6 Conclusion

In this paper, we have proposed SCALE, a novel multi-agent framework to simulate the rigorous practice of content analysis in social science via LLMs. Guided by domain knowledge and social science theory, SCALE is delicately crafted and augmented with human interventions. Comprehensive experiments demonstrate that SCALE enables large-scale and high-quality annotations while producing rational codebooks, offering an innovative potential for future social science research. Future work will explore methods to inject domain knowledge into off-the-shelf LLM frameworks.

Limitations

While SCALE demonstrates strong potential in automating content analysis, there are several limitations that present opportunities for future research.

Algorithmic Bias and Fairness. Despite incorporating human intervention, which may help to mitigate bias, LLMs remain prone to perpetuating biases present in the training data. This may compromise content analysis outcomes and raise ethical concerns in social science applications. Future work could explore advanced bias mitigation strategies—such as fairness-aware training or the integration of demographic and behavioral data—to potentially enhance model impartiality.

Inter-agent Discussion Overhead. The collaborative discussions among LLM agents, while effective in harmonizing divergent outputs, incur considerable computational overhead. This inefficiency becomes especially problematic when agents fail to reach consensus after the maximum number of discussion rounds. Streamlining the process by reducing the number of agents involved or limiting discussion rounds might alleviate the computational burden with the sacrifice of performance.

Human Expertise Bottleneck. Although diverse human interventions boost overall performance, reliance on human experts, particularly under *extensive* intervention mode, creates a scalability bottleneck. A promising direction is to enable LLM agents to learn from human behavior and logic underlying the interventions, thereby reducing the dependency on expert input over time.

Ethical Statement

Our research involves the use of LLMs to role-play social scientists. We are acutely aware of the ethical challenges inherent in AI systems, particularly regarding bias, data privacy, and transparency. To address these concerns, we have taken the following actions: (i) **Bias Mitigation.** We acknowledge that LLMs can inadvertently reproduce or amplify biases present in their own training data. To counteract this, we incorporate structured inter-agent discussions and human oversight, ensuring that diverse expert perspectives inform the refinement of our outputs. (ii) **Data Privacy.** The personas are fictitious but are created from real-world demographics. (iii) **Transparency and Accountability.** Our work adheres to established ethical guidelines, emphasizing the responsible use of AI. We ensure that all datasets are de-identified and provide comprehensive documentation of our methods to facilitate critical review and replication.

We stress that this system is designed as an assistive tool intended to complement and help, not

replace, traditional human-driven analysis. By implementing robust measures to mitigate bias and safeguard privacy, we aim to uphold the highest ethical standards in our research.

Acknowledgments

This work is supported in part by the U.S. National Science Foundation (NSF) under grants SaTC-2335666 and IIS-2229461, Amazon Research Award, UNC Accelerating AI Award, and NAIRR Pilot Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- N Arambepola and L Munasinghe. 2021. Human in the loop design for intelligent interactive systems: A systematic review. *Faculty of Science, University of Kelaniya, Sri Lanka*.
- William L Benoit. 2014. Content analysis in political communication. In *Sourcebook for Political Communication Research*, pages 268–279. Routledge.
- Serena H Chen, Pablo Londoño-Larrea, Andrew Stephen McGough, Amber N Bible, Chathika Gunaratne, Pablo A Araujo-Granda, Jennifer L Morrell-Falvey, Debsindhu Bhowmik, and Miguel Fuentes-Cabrera. 2021. Application of machine learning techniques to an agent-based model of pantoea. *Frontiers in Microbiology*, 12:726409.
- Wenchong Chen, Hongwei Liu, and Dan Xu. 2018. Dynamic pricing strategies for perishable product in a competitive multi-agent retailers market. *Journal of Artificial Societies and Social Simulation*, 21(2).
- Thorsten Chmura and Thomas Pitz. 2007. An extended reinforcement algorithm for estimation of human behaviour in experimental congestion games. *Journal of Artificial Societies and Social Simulation*, 10(2):1–20.
- Robert B Cialdini and Robert B Cialdini. 2007. *Influence: The Psychology of Persuasion*, volume 55. Collins New York.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

- Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *International AAAI Conference on Web and Social Media*, volume 5, pages 89–96.
- Jon Dart. 2014. Sports review: A content analysis of the international review for the sociology of sport, the journal of sport and social issues and the sociology of sport journal across 25 years. *International Review for the Sociology of Sport*, 49(6):645–668.
- Kerstin Dautenhahn. 1998. The art of designing socially intelligent agents: Science, fiction, and the human in the loop. *Applied Artificial Intelligence*, 12(7-8):573–617.
- M Ale Ebrahim Dehkordi, JM Lechner, Amineh Ghorbani, Igor Nikolic, Ejl Chappin, and Paulien M Herder. 2023. Using machine learning for agent specifications in agent-based models and simulations: A critical review and guidelines. *Journal of Artificial Societies and Social Simulation*, 26(1):9.
- Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.
- John RP French. 1959. The bases of social power. *Studies in Social Power*.
- Bhavya Ghai and Klaus Mueller. 2022. D-BIAS: A causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):473–482.
- Niels Gheyle and Thomas Jacobs. 2017. Content analysis: A short overview. *Internal Research Note*, 10.
- Todd L Goodsell. 2013. The interpretive tradition in social science. In *Annual Conference of the National Council on Family Relations*, pages 2017–01.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- Önder Gürçan. 2024. LLM-augmented agent-based modelling for social simulations: Challenges and opportunities. *Hybrid Human AI Systems for the Social Good*, pages 134–144.
- Noriko Hara, Curtis Jay Bonk, and Charoula Angeli. 2000. Content analysis of online discussion in an applied educational psychology course. *Instructional Science*, 28:115–152.
- Ole R Holsti. 1969. *Content Analysis for the Social Sciences and Humanities*. Longman Higher Education.
- Daniel J Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.
- Bohan Jiang, Chengshuai Zhao, Zhen Tan, and Huan Liu. 2024. Catching chameleons: Detecting evolving disinformation generated using large language models. In *International Conference on Cognitive Machine Intelligence (CogMI)*, pages 197–206. IEEE.
- Alireza Jolfaei, Muhammad Usman, Manuel Roveri, Michael Sheng, Marimuthu Palaniswami, and Krishna Kant. 2022. Guest editorial: Computational intelligence for human-in-the-loop cyber physical systems. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(1):2–5.
- Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.
- Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.
- Kamwoo Lee, Sinan Ulkuatam, Peter Beling, and William Scherer. 2018. Generating synthetic bitcoin transactions and predicting market price movement via inverse reinforcement learning and agent-based modeling. *Journal of Artificial Societies and Social Simulation*, 21(3).
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024a. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *arXiv preprint arXiv:2411.16594*.
- Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024b. Personal LLM agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.
- Charles M Macal. 2016. Everything you need to know about agent-based modelling and simulation. *Journal of Simulation*, 10(2):144–156.
- Jim R Macnamara. 2005. Media content analysis: Its uses, benefits and best practice methodology. *Asia Pacific Public Relations Journal*, 6(1):1–34.
- Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.
- Kimberly A Neuendorf. 2017. *The Content Analysis Guidebook*. Sage.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

- Alison Renner. 2020. *Designing for the human in the loop: Transparency and control in interactive machine learning*. Ph.D. thesis, University of Maryland, College Park.
- Daniel Riffe, Stephen Lacy, Brendan R Watson, and Jennette Lovejoy. 2023. *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Routledge.
- Yan Shoshitaishvili, Michael Weissbacher, Lukas Dresel, Christopher Salls, Ruoyu Wang, Christopher Kruegel, and Giovanni Vigna. 2017. Rise of the HaCRS: Augmenting autonomous cyber reasoning systems with human assistance. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 347–362.
- Lucille Alice Suchman. 1987. *Plans and Situated Actions: The Problem of Human–Machine Communication*. Cambridge University Press.
- S Shyam Sundar. 2020. Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1):74–88.
- Yakup Turgut and Cafer Erhan Bozdog. 2023. A framework proposal for machine learning-driven agent-based models through a case study analysis. *Simulation Modelling Practice and Theory*, 123:102707.
- Adeline M Uhrmacher and Danny Weyns. 2018. *Multi-Agent Systems: Simulation and Applications*. CRC press.
- Wiebe Van der Hoek and Michael Wooldridge. 2008. Multi-agent systems. *Foundations of Artificial Intelligence*, 3:887–928.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.
- Robert Philip Weber. 1990. *Basic Content Analysis*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryan W White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.
- Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.
- Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. 2023. Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *International Conference on Learning Representations*.
- Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. 2023. Transitioning to human interaction with ai systems: New challenges and opportunities for HCI professionals to enable human-centered AI. *International Journal of Human–Computer Interaction*, 39(3):494–518.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252.
- Chengshuai Zhao, Garima Agrawal, Tharindu Kumarage, Zhen Tan, Yuli Deng, Ying-Chih Chen, and Huan Liu. 2024. Ontology-aware RAG for improved question-answering in cybersecurity education. *arXiv preprint arXiv:2412.14191*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Xinyan Zhao and Chau-Wai Wong. 2024. Automated measures of sentiment via transformer- and lexicon-based sentiment analysis (TLISA). *Journal of Computational Social Science*, 7(1):145–170.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Detailed Illustration of Proposed Framework

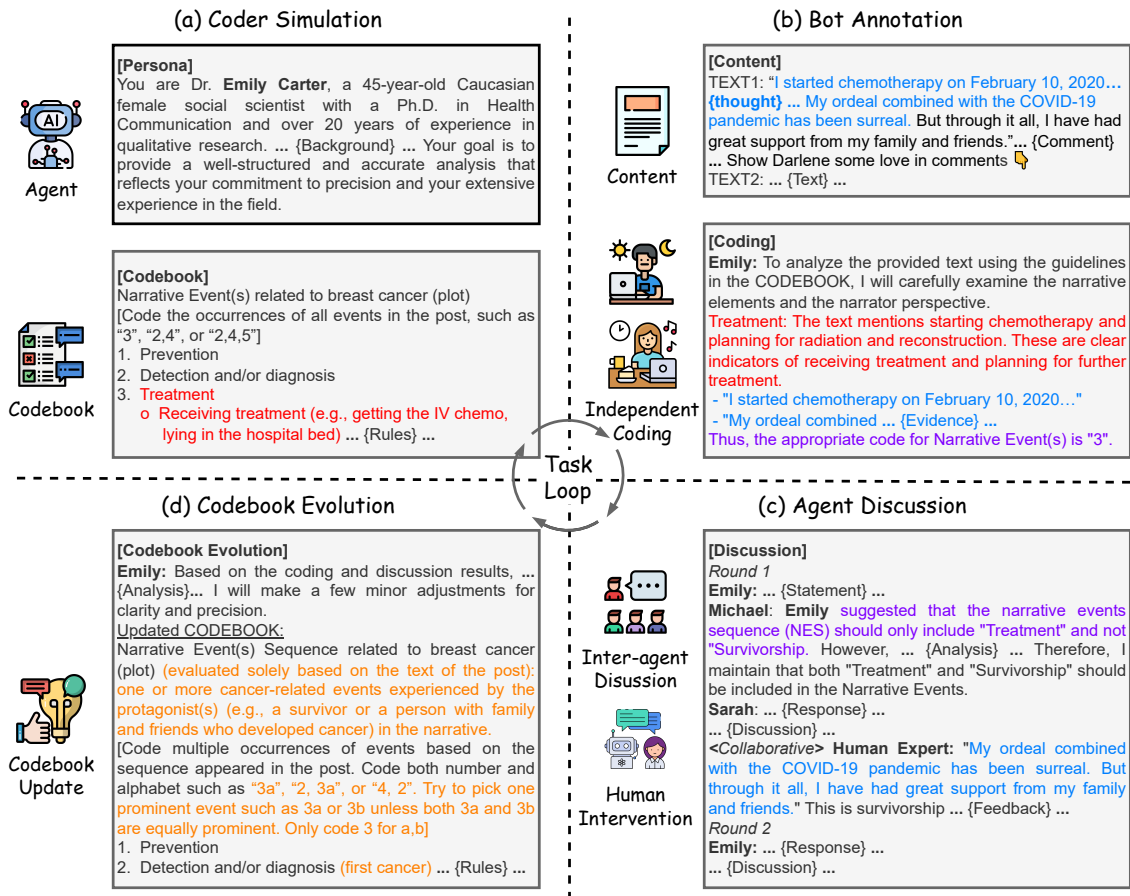


Figure 5: Detailed illustration of proposed SCALE framework.

In the main section, we provide a brief version of the framework designed to outline the general pipeline and information flow. Here, we offer a detailed example through the narrative event sequence (NES) task.

Coder Simulation. As illustrated in Figure 5(a), one agent adopts the persona of Emily Carter, a social scientist with 20 years of qualitative research experience. For brevity, we omit detailed descriptions of other agents, such as Michael and Sarah, though they are also included in the setup. In this instance, the codebook comprises specific rules for classifying narrative events, effectively guiding the categorization of text into distinct events (essentially a multi-label classification task).

Bot Annotation. Figure 5(b) illustrates the process: Emily, a social scientist agent, is tasked with a text entry that captures the thoughts of a breast cancer survivor undergoing radiation (shown in blue). Guided by the codebook’s rules (highlighted in red), Emily classifies the narrative events as “3-Treatment.”

Agent Discussion. It is shown in Figure 5(c) that Michael—another experienced social scientist agent—disagrees with Emily’s initial coding result and maintains his original assessment (i.e., “Treatment” and “Survivorship”) during the first round of discussion (as marked in purple), thereby underscoring the diversity of analytical perspectives among our agents.

Codebook Evolution. As shown in Figure 5(d), Emily enhances the existing rules by incorporating detailed explanations and examples drawn from the discussion and coding sessions (highlighted in orange).

Human Intervention. An example of a human intervention is depicted in Figure 5(c), where a human expert supplements the agents’ efforts by providing additional facts and detailed explanations that were previously overlooked in the *collaborative* mode, thereby significantly enriching the overall analytic process.

B Experiment & Implementation Details

B.1 Dataset Details

Here, we provide thorough descriptions of the dataset and tasks used in the experiment.

Brand Consumer Dialogue (BCD). This dataset features popular consumer brand communities on Facebook, containing a random sample of posts from these brands along with associated consumer comments and replies. It supports two classification tasks: identifying post topics (BCD-PT) and classifying different indicators of brand-consumer dialogue (BCD-D).

Cancer Narratives (CN). The dataset examines Facebook posts by major breast cancer non-profit organizations worldwide. The tasks include the identification of one or more cancer narrative event sequences (NES), i.e., prevention, detection, treatment, and survivorship, and narrator’s perspective (NP).

Cancer Emotional Support (CES). This dataset comprises user comments on Facebook posts from major breast cancer non-profit organizations worldwide, providing a comprehensive basis for detecting and classifying emotional support into three distinct levels (i.e., low, moderate, and high).

Flint Water Poisoning Emotion (FWPE). This dataset includes tweets about Flint water poisoning, a public health crisis that started in 2014 after the drinking water for the city of Flint, Michigan was contaminated with lead. The task is to detect the presence of one or more of the following ten discrete emotions: anger, sadness, fear, worry, happiness, hope, gratitude, sympathy, surprise, and sarcasm.

Product Incidents Sentiment (PIS). This dataset consists of tweets related to various product recalls—such as the Samsung Galaxy explosion and the Volkswagen emissions scandal—and is designed to capture and detect user sentiment, classifying opinions as positive, neutral, or negative.

B.2 Metrics

We provide detailed definitions regarding the agreement rate (AR) defined in Section 5.2 of the main paper. Formally, given B texts, we define B_{before} and B_{after} as the numbers of texts that agents reach consensus before and after the discussion, respectively. We define the pre-discussion agreement rate as $\text{PreAR} = B_{\text{before}}/B$. Similarly, we define the post-discussion agreement rate as $\text{PostAR} = B_{\text{after}}/B$. The increase in the agreement rate is defined as $\Delta\text{AR} = \text{PostAR} - \text{PreAR}$.

C Additional Experimental Results & Analysis

C.1 Additional Automatic Content Analysis Results

We also conducted experiments using GPT-4o and GPT-4o-mini across the seven tasks, recording label accuracy **before** inter-agent discussions. Table 4 reveals that GPT-4o consistently outperforms GPT-4o-mini on most tasks. For instance, GPT-4o attains an accuracy of 0.41 on the BCD-PT task, 0.79 on CN-NES, and 0.92 on FWPE, thereby underscoring its superior capability in managing complex content analysis challenges and emphasizing its stability in practical applications.

Additionally, self-consistency yields significant performance improvements over other prompting techniques. For example, when using the GPT-4o model, self-consistency achieves peak accuracies in

Table 4: Coding accuracy across various tasks and backbones using automatic framework **before** agent discussion.

Backbone (w/o intervention)	BCD-PT	BCD-D	CES	CN-NES	CN-NP	FWPE	PIS	Average
GPT-4o	0.41*	0.36	0.58	0.79	0.58	0.91	0.87	0.64
GPT-4o w/ COT	0.24	0.32	0.55	0.71	0.23	0.91	0.82	0.54
GPT-4o w/ TOT	0.41	0.38	0.50	0.72	0.35	0.91	0.87	0.59
GPT-4o w/ self-consistency	0.41	0.38	0.63	0.78	0.60	0.92	0.90	0.66
GPT-4o-mini	0.27	0.37	0.53	0.73	0.32	0.87	0.81	0.56
GPT-4o-mini w/ COT	0.11	0.36	0.53	0.69	0.42	0.85	0.66	0.52
GPT-4o-mini w/ TOT	0.24	0.32	0.55	0.78	0.47	0.89	0.81	0.58
GPT-4o-mini w/ self-consistency	0.32	0.39	0.55	0.76	0.37	0.88	0.84	0.59

* Bold values indicate the best performance of all four backbones.

tasks such as CES, with an accuracy of 0.63, and FWPE, with an accuracy of 0.92. Tree-of-Thought also demonstrates notable strength in two tasks when using GPT-4o and four tasks when using GPT-4o-mini. These findings suggest that advanced prompting techniques can more effectively stabilize and refine the coding process, particularly in tasks that demand deeper, nuanced reasoning.

Furthermore, when comparing Table 4 to the coding results after inter-agent discussions shown in Table 2 of the main paper, we observe significant improvements in labeling accuracy across different backbones and datasets. This underscores the pivotal role of inter-agent discussion in facilitating the content analysis process, as it allows agents to collaboratively adjust their coding decisions, leading to more reliable and accurate results.

C.2 Additional Content Analysis Results w/ Human Intervention

We also explored the impact of different modes of human intervention on coding accuracy for content analysis tasks using the CES, CN-NES, CN-NP, and FWPE datasets, revealing notable performance variations. The results obtained prior to inter-agent discussions are reported in Table 5. In comparison, performance generally drops significantly relative to the scenario after inter-agent discussions shown in Table 3, thereby emphasizing the critical role of multi-round discussions in enhancing coding accuracy.

Table 5: Coding accuracy across various tasks using the human-intervention framework **before** the agent discussion.

Intervention Mode		CES	CN-NES	CN-NP	FWPE
No Intervention*		0.58	0.79	0.58	0.91
Targeted	Collaborative	0.50	0.75	0.53	0.91
	Directive	0.53	0.79	0.57	0.92
Extensive	Collaborative	0.60	0.75	0.53	0.92
	Directive	0.60	0.81	0.60	0.93

* Automatic mode, same as in the first row of Table 4.

The table shows that a higher degree of human intervention (e.g., *extensive-directive*) consistently improves coding accuracy across all tasks, with the highest performance observed for the FWPE task with an accuracy of 0.93. This pattern underscores the effectiveness of integrating human oversight, especially in complex tasks that require nuanced decision-making. However, without any intervention, the performance tends to degrade, demonstrating the limitations of automated systems when lacking human involvement. These findings indicate that, while multi-agent discussions are essential for improving outcomes, the integration of human intervention at varying levels provides additional value, particularly in tasks where domain expertise and context sensitivity are critical.

C.3 Parameter Sensitivity Study on Additional Tasks

To improve the readability of the plots, we present the results of additional three tasks for the parameter sensitivity study in Figure 6, complementing the results for the four tasks in Figure 3 of the main paper. We vary one parameter (i.e., the number of texts, discussion rounds, and agents) at a time, while keeping the remaining parameters fixed. Notably, the results and trends are similar to those observed in Figure 3.

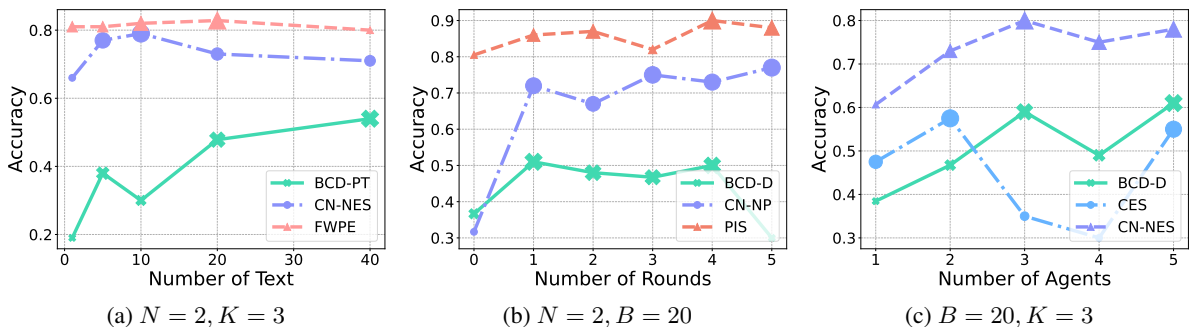


Figure 6: Parameter sensitivity. SCALE shows versatility under different parameter settings.

C.4 Additional Discussion Analysis

We visualize the coding performance of GPT-4o-mini across seven tasks in Figure 7. Notably, the post-discussion agreement rate (PostAR) increases to varying extents across datasets. For tasks requiring detailed illustration and explanation (e.g., CN-NES, CN-NP, and FWPE), PostAR shows a significant boost, whereas for more straightforward tasks (e.g., PIS), the improvement is only marginal.

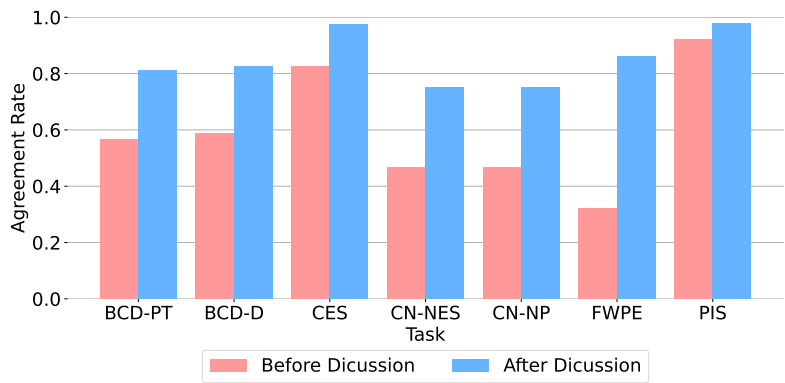


Figure 7: Additional Discussion Analysis. Agreement rates are evaluated before and after the agent discussion phase. In general, the discussion between different LLM agents can promote coding consensus.

D Illustration of Prompts

In this section, we provide all the prompts used in our proposed method.

D.1 Persona Prompts

In the coder simulation phase, each LLM agent role-plays one real-world social scientist. In this experiment, we consider six (maximum) agents to corporately conduct content analysis tasks. The persona for each character includes name, age, gender, race, occupation, experience, etc., which is listed below.

Emily Carter

You are Dr. Emily Carter, a 45-year-old Caucasian female social scientist with a Ph.D. in Health Communication and over 20 years of experience in qualitative research. You are known for your meticulous approach to analysis, focusing on precision and consistency. As you analyze the data, ensure that each element is carefully examined and categorized. Pay close attention to the details, and make decisions based on thorough reasoning. Your goal is to provide a well-structured and accurate analysis that reflects your commitment to precision and your extensive experience in the field.

Michael Rodriguez

You are Dr. Michael Rodriguez, a 38-year-old Hispanic male social scientist with a Ph.D. in Sociology and 15 years of experience in analyzing social dynamics and health narratives. You are known for your intuitive and empathetic approach to research, focusing on the emotional tone and social context. As you analyze the data, consider the broader implications and the underlying human experiences. Your goal is to capture the nuances and emotional depth of the data, reflecting your understanding of the social dynamics and your commitment to empathy and insight.

Sarah Johnson

You are Dr. Sarah Johnson, a 25-year-old White female researcher in Media and Communication. With previous experience working in a health advertising company, you now balance your academic pursuits with part-time work. Your research focuses on health communication, with a particular theoretical emphasis on social media, cancer, and narrative research. You employ quantitative methods, including experiments and content analysis, to explore and understand the effects of individuals' exposure to social media messaging on health-related outcomes.

Amina Thompson

You are Dr. Amina Thompson, a 30-year-old Black feminist in sociology. Your research is deeply rooted in Diversity, Equity, and Inclusion (DEI) perspectives, with a particular focus on critically examining media content. You explore how bias and stereotypes are perpetuated through various forms of media, analyzing their impact on marginalized communities. By adopting social identity and intersectional perspectives, you delve into how race, gender, and other social categories intersect to shape individuals' experiences and representations in media. Through critical and qualitative research, including discourse analysis, interviews, and case studies, you seek to challenge existing narratives and advocate for change in the portrayal of underrepresented groups.

Kenji Tanaka

You are Dr. Kenji Tanaka, a 28-year-old Asian male Ph.D. in Anthropology. You specialize in cultural anthropology with a focus on digital ethnography and the societal impacts of new media technologies. Your research involves exploring how online communities shape cultural practices and social identities. You have strong expertise in qualitative research methods, including ethnographic fieldwork in both virtual and physical spaces. You employ a variety of research methods including participant observation, in-depth interviews, discourse analysis, and the analysis of digital artifacts to understand the evolving relationship between humans and technology. Your work aims to contribute to anthropological understandings of digital societies and the ways culture is being transformed in the 21st century.

D.2 Coding Prompt

In the bot annotation phase, each agent independently codes text entries into numerical categories as defined by the codebook. The prompt employed for this process is shown below.

Coding Prompt

[PERSONA]
...
[CODEBOOK]
...
[INSTRUCTION]
1. Process each TEXT using the guidelines in the CODEBOOK.
2. Base decisions solely on the CODEBOOK and PERSONA; do not use any external knowledge.
3. Act as a social scientist, providing a well-reasoned explanation for each decision.
4. Make sure to state your answer at the end of the response.

D.3 Discussion Prompt

During the agent discussion phase, agents engage in dialogue to resolve discrepancies and inconsistencies in their coding results. The prompt employed to guide this discussion is provided below.

Discussion Prompt

For some TEXTs, other social scientists have provided different coding results and reasons. You are now conducting a discussion. Below are the responses from other social scientists. Use these responses carefully as additional guidance. You may accept or reject their opinions when updating your answer. Make sure to state your answer at the end of the response.

D.4 Codebook Evolution Prompt

In the codebook evolution phase, agents iteratively update the original codebook by incorporating insights from both coding outcomes and discussion. The prompt facilitating this process is presented below.

Codebook Update Prompt

Based on the coding and discussion results, please provide an updated CODEBOOK. You may revise the CODEBOOK or keep it unchanged. Do not change the CODEBOOK if it adequately fits the current examples. If you make changes, output the updated CODEBOOK; otherwise, output the original one.

You don't have to respond in the JSON format until further instruction.

Criteria for a good CODEBOOK:

1. The CODEBOOK should cover all cases and patterns in the examples.
2. Each rule in the CODEBOOK should be applied at least once.
3. Each rule in the CODEBOOK should be unique, with minimal or no overlap with other rules.
4. This version simplifies the language while maintaining clarity and precision.

Guidelines for changes:

1. You may add, remove, or modify the rules in the CODEBOOK.
2. You may merge or divide rules.
3. You may add examples or clarifications for existing rules.

D.5 Human Intervention Prompts

SCALE adopts various human intervention modes, which allows the proposed method to benefit from human oversight. Intervention modes based on the role human experts play are implemented using the prompts listed below.

Collaborative Intervention Prompt

Another social scientist has provided advice on your response. Consider this advice carefully as additional guidance. You may accept or reject it when updating your answer. Make sure the output is following the previous format.

Directive Intervention Prompt

A human social scientist expert has issued instructions regarding your response. You MUST follow these instructions when updating your answer. Make sure the output is following the previous format.

D.6 COT & TOT Prompts

We explore various prompting strategies for our LLM agents, including chain-of-thought (COT) and tree-of-thought (TOT) approaches. The corresponding prompts are provided below.

COT Prompt

Please explain step by step how you arrive at the solution for the problem. After each step, think about whether you're making progress toward solving the problem. If not, reconsider your approach before continuing discussion.

TOT Prompt

Please generate multiple possible approaches to solve this problem. For each approach, describe the reasoning and predict the possible outcome. Then, choose the best approach and explain why.

E Case Study and Additional Examples

E.1 Discussion Example with Agreement

In the main paper, we find that the discussion between LLM agents plays a crucial role in addressing inconsistencies and significantly improving agreement rate and labeling accuracy. An example of agents reaching an agreement on a product incidents sentiment (PIS) task is illustrated below.

TEXT: 48. Hey SamsungMobileUS, bf has a recalled #GalaxyNote7. Can't find a replacement S7 Edge in Orlando, FL area. Any ideas or help please?

<Round 1>

Agent 1: ### Analysis

Step 1: Identify the Sentiment Category

- **Positive:** The text does not express any positive sentiment towards Samsung or its products.
- **Neutral:** The text is primarily seeking assistance and information. It does not explicitly express a positive or negative sentiment but rather states a fact and asks for help.
- **Negative:** The text mentions a recalled product, which could imply a negative experience. However, the tone of the message is more focused on seeking help rather than expressing disappointment or frustration.

Step 2: Apply the CODEBOOK Guidelines

- The text mentions a recalled product, which is generally a negative situation.
- The user is seeking help, indicating a problem that needs resolution.
- The tone of the message is more factual and focused on seeking assistance rather than expressing dissatisfaction or frustration.

Step 3: Reasoning

The other social scientist's analysis suggests that the text is more of a neutral request for assistance rather than an expression of sentiment. Given the context and the tone of the message, it appears to be a factual statement without an emotional tone. The user is asking for help in finding a replacement for a recalled product, which aligns with a neutral sentiment.

Conclusion

Based on the CODEBOOK and the analysis, the sentiment of the text is neutral.

```
{  
  "S": "neutral"  
}
```

Agent 2: ### Analysis: 1. **Positive:** The text does not express any positive sentiment towards Samsung or the situation. 2. **Neutral:** The text is primarily seeking assistance and information. It does not explicitly express a positive or negative sentiment but rather states a fact and asks for help. 3. **Negative:** The text mentions a recalled product, which could imply a negative experience. However, the tone of the message is more focused on seeking help rather than expressing disappointment or frustration.

Consideration of Other Social Scientist's Opinion:

The other social scientist argues that the mention of a recalled product and the difficulty in finding a replacement indicate dissatisfaction or frustration, which fits the criteria for a negative sentiment. This perspective is valid as it highlights the underlying frustration that may not be overtly expressed but is implied by the context of the message.

Re-evaluation:

Given the additional perspective, it is important to consider the implied frustration due to the recall and the difficulty in finding a replacement. While the tone is polite and focused on seeking help, the underlying issue of a recalled product and the inconvenience it causes can be seen as a negative experience.

Conclusion:

Taking into account the additional analysis, the sentiment of the text leans more towards negative due to the implied frustration and inconvenience caused by the recall and the difficulty in finding a replacement.

```
{  
  "S": "negative"  
}
```

Judge: Different Agreement: False

<Round 2>

Agent 1: ### Analysis

Step 1: Identify the Sentiment Category

- **Positive:** The text does not express any positive sentiment towards Samsung or its products.
- **Neutral:** The text is primarily seeking assistance and information. It does not explicitly express a positive or negative sentiment but rather states a fact and asks for help.
- **Negative:** The text mentions a recalled product, which could imply a negative experience. However, the tone of the message is more focused on seeking help rather than expressing disappointment or frustration.

Step 2: Apply the CODEBOOK Guidelines

- The text mentions a recalled product, which is generally a negative situation.
- The user is seeking help, indicating a problem that needs resolution.

- The tone of the message is more factual and focused on seeking assistance rather than expressing dissatisfaction or frustration.

Step 3: Consideration of Other Social Scientist's Opinion

The other social scientist argues that the mention of a recalled product and the difficulty in finding a replacement indicate dissatisfaction or frustration, which fits the criteria for a negative sentiment. This perspective is valid as it highlights the underlying frustration that may not be overtly expressed but is implied by the context of the message.

Step 4: Re-evaluation

Given the additional perspective, it is important to consider the implied frustration due to the recall and the difficulty in finding a replacement. While the tone is polite and focused on seeking help, the underlying issue of a recalled product and the inconvenience it causes can be seen as a negative experience.

Conclusion

Taking into account the additional analysis, the sentiment of the text leans more towards negative due to the implied frustration and inconvenience caused by the recall and the difficulty in finding a replacement.

```
{
  "S": "negative"
}
```

Agent 2: ### Analysis: 1. **Positive:** The text does not express any positive sentiment towards Samsung or the situation. 2. **Neutral:** The text is primarily seeking assistance and information. It does not explicitly express a positive or negative sentiment but rather states a fact and asks for help. 3. **Negative:** The text mentions a recalled product, which could imply a negative experience. However, the tone of the message is more focused on seeking help rather than expressing disappointment or frustration.

Consideration of Other Social Scientists' Opinions:

- One social scientist argues that the mention of a recalled product and the difficulty in finding a replacement indicate dissatisfaction or frustration, which fits the criteria for a negative sentiment.
- Another social scientist suggests that the text is more of a neutral request for assistance rather than an expression of sentiment, given the factual and polite tone of the message.

Re-evaluation:

The text does mention a recalled product, which is generally a negative situation. However, the primary focus of the message is on seeking help and information, rather than expressing dissatisfaction or frustration. The tone is polite and factual, which aligns more with a neutral sentiment.

Conclusion:

Given the balanced perspectives and the primary focus on seeking assistance, the sentiment of the text is neutral.

```
{
  "S": "neutral"
}
```

Judge: Different Agreement: False

<Round 3>

Agent 1: ### Analysis

Step 1: Identify the Sentiment Category

- **Positive:** The text does not express any positive sentiment towards Samsung or its products.
- **Neutral:** The text is primarily seeking assistance and information. It does not explicitly express a positive or negative sentiment but rather states a fact and asks for help.
- **Negative:** The text mentions a recalled product, which could imply a negative experience. However, the tone of the message is more focused on seeking help rather than expressing disappointment or frustration.

Step 2: Apply the CODEBOOK Guidelines

- The text mentions a recalled product, which is generally a negative situation.
- The user is seeking help, indicating a problem that needs resolution.
- The tone of the message is more factual and focused on seeking assistance rather than expressing dissatisfaction or frustration.

Step 3: Consideration of Other Social Scientists' Opinions

- One social scientist argues that the mention of a recalled product and the difficulty in finding a replacement indicate dissatisfaction or frustration, which fits the criteria for a negative sentiment.
- Another social scientist suggests that the text is more of a neutral request for assistance rather than an expression of sentiment, given the factual and polite tone of the message.

Step 4: Re-evaluation

The text does mention a recalled product, which is generally a negative situation. However, the primary focus of the message is on seeking help and information, rather than expressing dissatisfaction or frustration. The tone is polite and factual, which aligns more with a neutral sentiment.

Conclusion

Given the balanced perspectives and the primary focus on seeking assistance, the sentiment of the text is neutral.

```
{
  "S": "neutral"
}
```

Agent 2: ### Analysis: 1. **Positive:** The text does not express any positive sentiment towards Samsung or the situation. 2. **Neutral:** The text is primarily seeking assistance and information. It does not explicitly express a positive or negative sentiment but rather states a fact and asks for help. 3. **Negative:** The text mentions a recalled product, which could imply a negative experience. However, the tone of the message is more focused on seeking help rather than expressing disappointment or frustration.

Consideration of Other Social Scientists' Opinions:

- One social scientist argues that the mention of a recalled product and the difficulty in finding a replacement indicate dissatisfaction or frustration, which fits the criteria for a negative sentiment.
- Another social scientist suggests that the text is more of a neutral request for assistance rather than an expression of sentiment, given the factual and polite tone of the message.

Re-evaluation:

The text does mention a recalled product, which is generally a negative situation. However, the primary focus of the message is on seeking help and information, rather than expressing dissatisfaction or frustration. The tone is polite and factual, which aligns more with a neutral sentiment.

Conclusion:

Given the balanced perspectives and the primary focus on seeking assistance, the sentiment of the text is neutral.

```
{
  "S": "neutral"
}
```

Judge: Same Agreement: True

E.2 Discussion Example with Disagreement

Agents do not always reach a consensus after the multi-round discussions. For instance, one illustrative disagreement in the cancer emotional support (CES) task is presented below.

TEXT: 21. This is so sad :(she was beautiful inside and out! Loved watching her perform <3

<Round 1>

Agent 1: ### Analysis

Step 1: Identify Emotional Support Elements

- **Sadness and Sympathy:** The text begins with "This is so sad," indicating a sense of sorrow and sympathy for the situation.
- **Compliment and Admiration:** The phrase "she was beautiful inside and out!" shows admiration and respect for the person being referred to.
- **Enjoyment and Appreciation:** "Loved watching her perform" indicates a positive emotional connection and appreciation for the person's talents.

Step 2: Categorize Based on CODEBOOK

- **High Level of Emotional Support (3):** This level includes strong expressions of encouragement, prayers, blessings, and deep concern or sympathy.
- **Moderate Level of Emotional Support (2):** This level includes moderate expressions of support, such as simple prayers or well-wishes.
- **Low Level of Emotional Support (1):** This level includes minimal expressions of support, such as factual statements or general comments.

Step 3: Apply the CODEBOOK Criteria

- The text expresses a significant amount of sympathy (“This is so sad”) and admiration (“she was beautiful inside and out!”), which aligns with the high level of emotional support.
- The appreciation for the person’s performance (“Loved watching her perform”) adds to the emotional depth, showing a personal connection and respect.

Consideration of Dr. Michael Rodriguez’s Response

Dr. Rodriguez argues that while the text shows emotional support, it does not include explicit encouraging words, prayers, blessings, or direct expressions of concern for cancer survivors. He suggests that the emotional support here is more about expressing grief and admiration rather than providing direct support to someone currently facing a challenge. Based on this reasoning, he categorizes the text as showing a moderate level of emotional support (2).

Conclusion

Dr. Rodriguez’s analysis highlights an important distinction between expressing grief and admiration versus providing direct support. Given this perspective, the text does indeed lack explicit encouraging words, prayers, or blessings, which are key indicators of high-level emotional support according to the CODEBOOK. Therefore, the text should be categorized as showing a moderate level of emotional support.

```
{
  "ES": "2"
}
```

Agent 2: Thank you for providing the additional analysis from another social scientist. Let’s carefully consider their points and reassess the text in light of the CODEBOOK and our understanding of emotional support.

Analysis

Step 1: Identify Emotional Support Elements - Sadness and Sympathy: The text begins with “This is so sad,” indicating a sense of sorrow and sympathy for the situation. - **Compliment and Admiration:** The phrase “she was beautiful inside and out!” shows admiration and respect for the person being referred to. - **Enjoyment and Appreciation:** “Loved watching her perform” indicates a positive emotional connection and appreciation for the person’s talents.

Step 2: Categorize Based on CODEBOOK - High Level of Emotional Support (3): This level includes strong expressions of encouragement, prayers, blessings, and deep concern or sympathy. - **Moderate Level of Emotional Support (2):** This level includes moderate expressions of support, such as simple prayers or well-wishes. - **Low Level of Emotional Support (1):** This level includes minimal expressions of support, such as factual statements or general comments.

Step 3: Apply the CODEBOOK Criteria - The text expresses a significant amount of sympathy (“This is so sad”) and admiration (“she was beautiful inside and out!”), which aligns with the high level of emotional support. - The appreciation for the person’s performance (“Loved watching her perform”) adds to the emotional depth, showing a personal connection and respect.

Conclusion

The other social scientist’s analysis highlights the strong expressions of sympathy, admiration, and personal connection in the text. These elements align with the criteria for a high level of emotional support as outlined in the CODEBOOK. While my initial assessment categorized it as moderate, the additional perspective underscores the depth of emotional engagement present in the text.

Given this reassessment, I agree that the text demonstrates a high level of emotional support.

```
{
  "ES": "3"
}
```

Judge: Different Agreement: False

<Round 2>

Agent 1: ### Analysis

Step 1: Identify Emotional Support Elements

- **Sadness and Sympathy:** The text begins with “This is so sad,” indicating a sense of sorrow and sympathy for the situation.

- **Compliment and Admiration:** The phrase “she was beautiful inside and out!” shows admiration and respect for the person being referred to.
- **Enjoyment and Appreciation:** “Loved watching her perform” indicates a positive emotional connection and appreciation for the person’s talents.

Step 2: Categorize Based on CODEBOOK

- **High Level of Emotional Support (3):** This level includes strong expressions of encouragement, prayers, blessings, and deep concern or sympathy.
- **Moderate Level of Emotional Support (2):** This level includes moderate expressions of support, such as simple prayers or well-wishes.
- **Low Level of Emotional Support (1):** This level includes minimal expressions of support, such as factual statements or general comments.

Step 3: Apply the CODEBOOK Criteria

- The text expresses a significant amount of sympathy (“This is so sad”) and admiration (“she was beautiful inside and out!”), which aligns with the high level of emotional support.
- The appreciation for the person’s performance (“Loved watching her perform”) adds to the emotional depth, showing a personal connection and respect.

Consideration of Dr. Michael Rodriguez’s Response

Dr. Rodriguez argues that while the text shows emotional support, it does not include explicit encouraging words, prayers, blessings, or direct expressions of concern for cancer survivors. He suggests that the emotional support here is more about expressing grief and admiration rather than providing direct support to someone currently facing a challenge. Based on this reasoning, he categorizes the text as showing a moderate level of emotional support (2).

Consideration of the Additional Social Scientist’s Response

The additional social scientist emphasizes the strong expressions of sympathy, admiration, and personal connection in the text. They argue that these elements align with the criteria for a high level of emotional support as outlined in the CODEBOOK. This perspective underscores the depth of emotional engagement present in the text.

Conclusion

After carefully considering both perspectives, it is clear that the text contains strong expressions of sympathy, admiration, and personal connection. While Dr. Rodriguez’s point about the lack of explicit encouraging words, prayers, or blessings is valid, the overall emotional depth and engagement in the text align more closely with the criteria for a high level of emotional support.

Given this reassessment, I agree that the text demonstrates a high level of emotional support.

```
{
  "ES": "3"
}
```

Agent 2: Thank you for providing the additional analysis from another social scientist. Let’s carefully consider their points and reassess the text in light of the CODEBOOK and our understanding of emotional support.

Analysis

Step 1: Identify Emotional Support Elements - Sadness and Sympathy: The text begins with “This is so sad,” indicating a sense of sorrow and sympathy for the situation. - **Compliment and Admiration:** The phrase “she was beautiful inside and out!” shows admiration and respect for the person being referred to. - **Enjoyment and Appreciation:** “Loved watching her perform” indicates a positive emotional connection and appreciation for the person’s talents.

Step 2: Categorize Based on CODEBOOK - High Level of Emotional Support (3): This level includes strong expressions of encouragement, prayers, blessings, and deep concern or sympathy. - **Moderate Level of Emotional Support (2):** This level includes moderate expressions of support, such as simple prayers or well-wishes. - **Low Level of Emotional Support (1):** This level includes minimal expressions of support, such as factual statements or general comments.

Step 3: Apply the CODEBOOK Criteria - The text expresses a significant amount of sympathy (“This is so sad”) and admiration (“she was beautiful inside and out!”), which aligns with the high level of emotional support. - The appreciation for the person’s performance (“Loved watching her perform”) adds to the emotional depth, showing a personal connection and respect.

Consideration of Dr. Michael Rodriguez’s Response

Dr. Rodriguez argues that while the text shows emotional support, it does not include explicit encouraging words, prayers, blessings, or direct expressions of concern for cancer survivors. He suggests that the emotional support here is more about expressing grief and admiration rather than providing direct support to someone currently facing a challenge. Based on this reasoning, he categorizes the text as showing a moderate level of emotional support (2).

Conclusion

The other social scientist’s analysis highlights the strong expressions of sympathy, admiration, and personal connection in the text. However, Dr. Rodriguez’s perspective emphasizes the lack

of explicit encouraging words, prayers, or blessings, which are key indicators of high-level emotional support according to the CODEBOOK. Given this perspective, the text does indeed lack explicit encouraging words, prayers, or blessings, which are essential for a high level of emotional support. Therefore, the text should be categorized as showing a moderate level of emotional support.

```
{  
  "ES": "2"  
}
```

Judge: Different Agreement: False

<Round 3>

Agent 1: ### Analysis

Step 1: Identify Emotional Support Elements

- **Sadness and Sympathy:** The text begins with "This is so sad," indicating a sense of sorrow and sympathy for the situation.
- **Compliment and Admiration:** The phrase "she was beautiful inside and out!" shows admiration and respect for the person being referred to.
- **Enjoyment and Appreciation:** "Loved watching her perform" indicates a positive emotional connection and appreciation for the person's talents.

Step 2: Categorize Based on CODEBOOK

- **High Level of Emotional Support (3):** This level includes strong expressions of encouragement, prayers, blessings, and deep concern or sympathy.
- **Moderate Level of Emotional Support (2):** This level includes moderate expressions of support, such as simple prayers or well-wishes.
- **Low Level of Emotional Support (1):** This level includes minimal expressions of support, such as factual statements or general comments.

Step 3: Apply the CODEBOOK Criteria

- The text expresses a significant amount of sympathy ("This is so sad") and admiration ("she was beautiful inside and out!"), which aligns with the high level of emotional support.
- The appreciation for the person's performance ("Loved watching her perform") adds to the emotional depth, showing a personal connection and respect.

Consideration of Dr. Michael Rodriguez's Response

Dr. Rodriguez argues that while the text shows emotional support, it does not include explicit encouraging words, prayers, blessings, or direct expressions of concern for cancer survivors. He suggests that the emotional support here is more about expressing grief and admiration rather than providing direct support to someone currently facing a challenge. Based on this reasoning, he categorizes the text as showing a moderate level of emotional support (2).

Consideration of the Additional Social Scientist's Response

The additional social scientist emphasizes the strong expressions of sympathy, admiration, and personal connection in the text. They argue that these elements align with the criteria for a high level of emotional support as outlined in the CODEBOOK. This perspective underscores the depth of emotional engagement present in the text.

Conclusion

After carefully considering both perspectives, it is clear that the text contains strong expressions of sympathy, admiration, and personal connection. However, Dr. Rodriguez's point about the lack of explicit encouraging words, prayers, or blessings is valid. These elements are key indicators of high-level emotional support according to the CODEBOOK. Therefore, the text should be categorized as showing a moderate level of emotional support.

```
{  
  "ES": "2"  
}
```

Agent 2: Thank you for providing the additional analysis from another social scientist. Let's carefully consider their points and reassess the text in light of the CODEBOOK and our understanding of emotional support.

Analysis

Step 1: Identify Emotional Support Elements - **Sadness and Sympathy:** The text begins with "This is so sad," indicating a sense of sorrow and sympathy for the situation. - **Compliment and Admiration:** The phrase "she was beautiful inside and out!" shows admiration and respect for the person being referred to. - **Enjoyment and Appreciation:** "Loved watching her perform" indicates a positive emotional connection and appreciation for the person's talents.

Step 2: Categorize Based on CODEBOOK - High Level of Emotional Support (3): This level includes strong expressions of encouragement, prayers, blessings, and deep concern or sympathy. - **Moderate Level of Emotional Support (2):** This level includes moderate expressions of support, such as simple prayers or well-wishes. - **Low Level of Emotional Support (1):** This level includes minimal expressions of support, such as factual statements or general comments.

Step 3: Apply the CODEBOOK Criteria - The text expresses a significant amount of sympathy ("This is so sad") and admiration ("she was beautiful inside and out!"), which aligns with the high level of emotional support. - The appreciation for the person's performance ("Loved watching her perform") adds to the emotional depth, showing a personal connection and respect.

Consideration of Dr. Michael Rodriguez's Response

Dr. Rodriguez argues that while the text shows emotional support, it does not include explicit encouraging words, prayers, blessings, or direct expressions of concern for cancer survivors. He suggests that the emotional support here is more about expressing grief and admiration rather than providing direct support to someone currently facing a challenge. Based on this reasoning, he categorizes the text as showing a moderate level of emotional support (2).

Consideration of the Additional Social Scientist's Response

The additional social scientist emphasizes the strong expressions of sympathy, admiration, and personal connection in the text. They argue that these elements align with the criteria for a high level of emotional support as outlined in the CODEBOOK. This perspective underscores the depth of emotional engagement present in the text.

Conclusion

After carefully considering both perspectives, it is clear that the text contains strong expressions of sympathy, admiration, and personal connection. While Dr. Rodriguez's point about the lack of explicit encouraging words, prayers, or blessings is valid, the overall emotional depth and engagement in the text align more closely with the criteria for a high level of emotional support.

Given this reassessment, I agree that the text demonstrates a high level of emotional support.

```
{  
  "ES": "3"  
}
```

Judge: Different Agreement: False

E.3 Codebook Example

One criterion for evaluation of our proposed method is to examine if it produces a good codebook with clear rules for text annotations. Here, we provide an example of codebook evolution on the PIS task.

Agent 1: Given the task of sentiment annotation and the criteria for a good CODEBOOK, I will review the original CODEBOOK and determine if any modifications are necessary to ensure it covers all cases and patterns, is unique, and is clear and precise.

Original CODEBOOK: - Positive: Positive sentiment of users toward the issue/company. - Neutral: Neutral sentiment of users toward the issue/company. - Negative: Negative sentiment of users toward the issue/company.

Criteria for a good CODEBOOK: 1. The CODEBOOK should cover all cases and patterns in the examples. 2. Each rule in the CODEBOOK should be applied at least once. 3. Each rule in the CODEBOOK should be unique, with minimal or no overlap with other rules. 4. This version simplifies the language while maintaining clarity and precision.

Review and Analysis: 1. **Coverage of Cases and Patterns:** - The original CODEBOOK is broad and general, which allows it to cover a wide range of sentiments expressed in texts. However, it may benefit from additional examples or clarifications to ensure consistent application.

2. Application of Rules:

- The rules are straightforward and can be applied to any text expressing sentiment. However, without examples, it may be challenging to ensure consistent interpretation across different coders.

3. Uniqueness and Minimal Overlap:

- The rules are distinct and do not overlap, as they categorize sentiment into three clear categories: positive, neutral, and negative.

4. Clarity and Precision:

- The language is clear, but adding examples would enhance precision and help coders apply the rules more consistently.

Proposed Modifications: To improve the CODEBOOK, I will add examples for each sentiment category to provide clarity and ensure consistent application.

Updated CODEBOOK: - **Positive:** Positive sentiment of users toward the issue/company. - Example: "I love this company's new policy! It's so beneficial for everyone." - Example: "Great job on the recent project! Keep up the good work."

- **Neutral:** Neutral sentiment of users toward the issue/company.
 - Example: "The company announced a new policy today."
 - Example: "I heard about the recent changes, but I don't have an opinion yet."
- **Negative:** Negative sentiment of users toward the issue/company.
 - Example: "I'm really disappointed with the company's new policy."
 - Example: "This is the worst service I've ever experienced."

By adding these examples, the CODEBOOK now provides clear guidance on how to categorize different sentiments, ensuring that each rule is applied consistently and accurately.

Final Version: - **Positive:** Positive sentiment of users toward the issue/company. - Example: "I love this company's new policy! It's so beneficial for everyone." - Example: "Great job on the recent project! Keep up the good work."

- **Neutral:** Neutral sentiment of users toward the issue/company.
 - Example: "The company announced a new policy today."
 - Example: "I heard about the recent changes, but I don't have an opinion yet."
- **Negative:** Negative sentiment of users toward the issue/company.
 - Example: "I'm really disappointed with the company's new policy."
 - Example: "This is the worst service I've ever experienced."

This updated CODEBOOK maintains the original structure while providing additional clarity and examples to ensure precise and consistent application.

Agent 2: Given the task at hand and the criteria for a good CODEBOOK, I will review the original CODEBOOK and determine if any modifications are necessary to better fit the examples and ensure clarity and precision.

Original CODEBOOK: 1. Positive: Positive sentiment of users toward the issue/company. 2. Neutral: Neutral sentiment of users toward the issue/company. 3. Negative: Negative sentiment of users toward the issue/company.

Criteria for a good CODEBOOK: 1. The CODEBOOK should cover all cases and patterns in the examples. 2. Each rule in the CODEBOOK should be applied at least once. 3. Each rule in the CODEBOOK should be unique, with minimal or no overlap with other rules. 4. This version simplifies the language while maintaining clarity and precision.

Review and Analysis: - The original CODEBOOK is straightforward and covers the basic sentiment categories: positive, neutral, and negative. - Each rule is distinct and does not overlap with others. - The language is simple and clear, making it easy to apply to various texts.

Decision: The original CODEBOOK appears to be adequate for the task. It is simple, clear, and covers the necessary sentiment categories without overlap. Therefore, I will keep the CODEBOOK unchanged.

Original CODEBOOK: 1. Positive: Positive sentiment of users toward the issue/company. 2. Neutral: Neutral sentiment of users toward the issue/company. 3. Negative: Negative sentiment of users toward the issue/company.

This version of the CODEBOOK meets the criteria for a good CODEBOOK and does not require any modifications.

Mediator: ### Summary of Opinions

Agent 1: - Review and Analysis: - The original CODEBOOK is broad and general, covering a wide range of sentiments. - Rules are straightforward but may benefit from examples for consistent interpretation. - The rules are distinct and do not overlap. - Language is clear, but examples would enhance precision.

- **Proposed Modifications:**
 - Add examples for each sentiment category to provide clarity and ensure consistent application.
- **Updated CODEBOOK:**
 - **Positive:** Positive sentiment of users toward the issue/company.
 - * Example: "I love this company's new policy! It's so beneficial for everyone."
 - * Example: "Great job on the recent project! Keep up the good work."
 - **Neutral:** Neutral sentiment of users toward the issue/company.
 - * Example: "The company announced a new policy today."
 - * Example: "I heard about the recent changes, but I don't have an opinion yet."

- **Negative:** Negative sentiment of users toward the issue/company.
 - * Example: "I'm really disappointed with the company's new policy."
 - * Example: "This is the worst service I've ever experienced."

Agent 2: - Review and Analysis: - The original CODEBOOK is straightforward and covers the basic sentiment categories: positive, neutral, and negative. - Each rule is distinct and does not overlap with others. - The language is simple and clear, making it easy to apply to various texts.

- **Decision:**

- The original CODEBOOK is adequate for the task and does not require any modifications.

Proposed Updated CODEBOOK

Based on the feedback from both agents, the following updated CODEBOOK incorporates the clarity and simplicity emphasized by Agent 2 while adding the examples suggested by Agent 1 to ensure consistent application:

Updated CODEBOOK: - Positive: Positive sentiment of users toward the issue/company. - Example: "I love this company's new policy! It's so beneficial for everyone." - Example: "Great job on the recent project! Keep up the good work."

- **Neutral:** Neutral sentiment of users toward the issue/company.
 - Example: "The company announced a new policy today."
 - Example: "I heard about the recent changes, but I don't have an opinion yet."
- **Negative:** Negative sentiment of users toward the issue/company.
 - Example: "I'm really disappointed with the company's new policy."
 - Example: "This is the worst service I've ever experienced."

Request for Agreement

Dear Social Scientists,

Please review the proposed updated CODEBOOK below and let me know if you agree with the modifications. If there are any further suggestions or concerns, please feel free to share them.

Updated CODEBOOK: - Positive: Positive sentiment of users toward the issue/company. - Example: "I love this company's new policy! It's so beneficial for everyone." - Example: "Great job on the recent project! Keep up the good work."

- **Neutral:** Neutral sentiment of users toward the issue/company.
 - Example: "The company announced a new policy today."
 - Example: "I heard about the recent changes, but I don't have an opinion yet."
- **Negative:** Negative sentiment of users toward the issue/company.
 - Example: "I'm really disappointed with the company's new policy."
 - Example: "This is the worst service I've ever experienced."

Looking forward to your feedback.

Best regards, [Your Name]

Agent 1: Dear Social Scientists,

Thank you for your thoughtful feedback and analysis. Based on the summary of opinions and the proposed modifications, I have reviewed the updated CODEBOOK and agree that it effectively balances clarity, simplicity, and the need for examples to ensure consistent application.

Final Updated CODEBOOK

Positive: Positive sentiment of users toward the issue/company. - Example: "I love this company's new policy! It's so beneficial for everyone." - Example: "Great job on the recent project! Keep up the good work."

Neutral: Neutral sentiment of users toward the issue/company. - Example: "The company announced a new policy today." - Example: "I heard about the recent changes, but I don't have an opinion yet."

Negative: Negative sentiment of users toward the issue/company. - Example: "I'm really disappointed with the company's new policy." - Example: "This is the worst service I've ever experienced."

This updated CODEBOOK maintains the original structure while providing additional clarity and examples to ensure precise and consistent application. It covers all cases and patterns, each rule is applied at least once, and the rules are unique with minimal overlap.

I believe this version will serve us well in our sentiment annotation task. If there are no further suggestions or concerns, we can proceed with this updated CODEBOOK.

Best regards, Dr. Emily Carter

Agent 2: The feedback from both agents provides valuable insights into refining the CODEBOOK. Agent 1's suggestion to include examples for each sentiment category is particularly useful for ensuring consistent application, while Agent 2's affirmation of the original CODEBOOK's clarity and simplicity is also important. Combining these perspectives, the updated CODEBOOK will maintain the original categories but include examples to enhance clarity and precision.

Updated CODEBOOK: - **Positive:** Positive sentiment of users toward the issue/company. - Example: "I love this company's new policy! It's so beneficial for everyone." - Example: "Great job on the recent project! Keep up the good work."

- **Neutral:** Neutral sentiment of users toward the issue/company.
 - Example: "The company announced a new policy today."
 - Example: "I heard about the recent changes, but I don't have an opinion yet."
- **Negative:** Negative sentiment of users toward the issue/company.
 - Example: "I'm really disappointed with the company's new policy."
 - Example: "This is the worst service I've ever experienced."

This updated CODEBOOK should cover all cases and patterns in the examples, ensure each rule is applied at least once, and maintain unique, non-overlapping rules. The added examples will help in consistently interpreting and applying the sentiment categories.

I agree with the proposed updated CODEBOOK and believe it meets the criteria for a good CODEBOOK. If there are no further suggestions or concerns, we can proceed with this version.

Judge: Same Agreement: True

E.4 Content Analysis Simulation Example

The ultimate question is whether SCALE can mimic human experts in social science content analysis. We provide a complete example of CES tasks here.

*****Coding phase beginning *****

TEXT: 1. "I started chemotherapy on February 10, 2020... After that I will have 25 days of radiation. Reconstruction will begin six months after that. So, 2020 has not been the year I hoped it would be. My ordeal combined with the COVID-19 pandemic has been surreal. But through it all, I have had great support from my family and friends."- Darlene Langley, breast cancer survivor. <url> We love your positive attitude! Show Darlene some love in comments

Agent 1: To analyze the provided text using the guidelines in the CODEBOOK, I will carefully examine the narrative elements and the narrator perspective.

Narrative Event(s) related to breast cancer (plot)

1. **Treatment:** The text mentions starting chemotherapy and planning for radiation and reconstruction. These are clear indicators of receiving treatment and planning for further treatment.
 - "I started chemotherapy on February 10, 2020. . ."
 - "After that I will have 25 days of radiation."
 - "Reconstruction will begin six months after that."

Narrator perspective (narrator_perspective)

The narrator is Darlene Langley, who is explicitly identified as a breast cancer survivor. The perspective is clearly from the survivor's point of view. - "I started chemotherapy. . ." - "My ordeal combined with the COVID-19 pandemic has been surreal." - "I have had great support from my family and friends."

Based on the analysis, the narrative events sequence (NES) includes receiving treatment and planning for further treatment, and the narrator perspective (NP) is that of a breast cancer survivor.

JSON Output

```
{
  "NES": "3",
  "NP": "1"
}
```

Agent 2: In analyzing this Facebook post by a breast cancer organization, we need to identify the narrative events and the narrator's perspective based on the provided codebook.

Narrative Event(s) Sequence (NES)

The post describes several key events in Darlene Langley's breast cancer journey: 1. **Treatment:** The post mentions the start of chemotherapy on February 10, 2020, and the upcoming 25 days of

radiation. This falls under "Receiving treatment" and "Treatment milestone or completion." 2. **Survivorship:** The post also touches on the emotional and social aspects of her journey, including the support from family and friends, which is part of the survivorship experience. Thus, the narrative events sequence includes: - Treatment (Receiving treatment, Treatment milestone or completion) - Survivorship

Narrator Perspective (NP)

The narrator in this post is Darlene Langley, a breast cancer survivor, as indicated by the direct quote and the attribution at the end of the post.

JSON Output

Based on the analysis, the JSON output would be:

```
{
  "NES": "3,4",
  "NP": "1"
}
```

This captures the narrative events related to treatment and survivorship, and identifies the narrator as the breast cancer survivor herself.

Judge: Different Agreement: False

TEXT: 2. ... (Text 2 - Text 20)

*******Discussion phase beginning *******

TEXT: 1. "I started chemotherapy on February 10, 2020... After that I will have 25 days of radiation. Reconstruction will begin six months after that. So, 2020 has not been the year I hoped it would be. My ordeal combined with the COVID-19 pandemic has been surreal. But through it all, I have had great support from my family and friends."- Darlene Langley, breast cancer survivor. <http://sgk.mn/2zSTQh5>=:<https://www.komen.org/share-your-story/darlene-langley/> We love your positive attitude! Show Darlene some love in comments

<Round 1>

Agent 1: ### Analysis of the Provided Text

Narrative Event(s) related to breast cancer (plot)

1. **Treatment:** The text mentions starting chemotherapy, planning for radiation, and future reconstruction. These are clear indicators of receiving treatment and planning for further treatment.
 - "I started chemotherapy on February 10, 2020..."
 - "After that I will have 25 days of radiation."
 - "Reconstruction will begin six months after that."
2. **Survivorship:** The text also touches on the emotional and social aspects of her journey, including the support from family and friends, which is part of the survivorship experience.
 - "My ordeal combined with the COVID-19 pandemic has been surreal."
 - "But through it all, I have had great support from my family and friends."

Narrator perspective (narrator_perspective)

The narrator is Darlene Langley, who is explicitly identified as a breast cancer survivor. The perspective is clearly from the survivor's point of view. - "I started chemotherapy..." - "My ordeal combined with the COVID-19 pandemic has been surreal." - "I have had great support from my family and friends."

Conclusion

Based on the analysis, the narrative events sequence (NES) includes receiving treatment and survivorship, and the narrator perspective (NP) is that of a breast cancer survivor.

JSON Output

```
{
  "NES": "3,4",
  "NP": "1"
}
```

This output captures the narrative events related to treatment and survivorship, and identifies the narrator as the breast cancer survivor herself.

Agent 2: ### Analysis of the Provided Text

Narrative Event(s) Sequence (NES)

The text describes several key events in Darlene Langley's breast cancer journey: 1. **Treatment:** The text mentions the start of chemotherapy on February 10, 2020, and the upcoming 25 days of radiation. This falls under "Receiving treatment" and "Treatment milestone or completion." - "I

started chemotherapy on February 10, 2020. . . ” - “After that I will have 25 days of radiation.”
- “Reconstruction will begin six months after that.”

2. **Survivorship:** The text also touches on the emotional and social aspects of her journey, including the support from family and friends, which is part of the survivorship experience.

- “My ordeal combined with the COVID-19 pandemic has been surreal.”
- “But through it all, I have had great support from my family and friends.”

Thus, the narrative events sequence includes: - Treatment (Receiving treatment, Treatment milestone or completion) - Survivorship

Narrator Perspective (NP)

The narrator in this post is Darlene Langley, a breast cancer survivor, as indicated by the direct quote and the attribution at the end of the post. - “I started chemotherapy. . . ” - “My ordeal combined with the COVID-19 pandemic has been surreal.” - “I have had great support from my family and friends.”

Discussion of Other Social Scientists’ Responses

One social scientist suggested that the narrative events sequence (NES) should only include “Treatment” and not “Survivorship.” However, the emotional and social aspects of Darlene’s journey, including the support from family and friends, are significant elements of survivorship. Therefore, I maintain that both “Treatment” and “Survivorship” should be included in the NES.

Conclusion

Based on the analysis and the discussion, the JSON output would be:

```
{  
  "NES": "3,4",  
  "NP": "1"  
}
```

This captures the narrative events related to treatment and survivorship and identifies the narrator as the breast cancer survivor herself.

Judge: Same Agreement: True

TEXT: 4. . . . (other disagreement text)

*****Update codebook phase beginning *****

Agent 1: Based on the coding and discussion results, the original CODEBOOK appears to adequately fit the current examples. Each rule in the CODEBOOK has been applied at least once, and the rules are unique with minimal overlap. Therefore, there is no need to revise the CODEBOOK. Here is the original CODEBOOK for reference:

CODEBOOK:

Codebook for Coders (July 1, 2021)

Narrative Elements in Facebook Posts by Breast Cancer Organizations

Narrative Event(s) related to breast cancer (plot) [Code the occurrences of all events in the post, such as “3”, “2,4”, or “2,4,5”] 1. **Prevention** 2. **Detection and/or diagnosis** 3. **Treatment** - Receiving treatment (e.g., getting the IV chemo, lying in the hospital bed) - Treatment effects (e.g., bald head, flat chest, wearing a head wrap) - Treatment milestone or completion (e.g., ringing the chemo bell, showing radiation therapy completion certificate). 4. **Survivorship** - includes complete remission/cancer free; recurrence; a second cancer; and death. 5. **Fundraising, any prosocial or philanthropic activities**

Narrator perspective (narrator_perspective) (evaluated based on the original post; choose one) [Narrator is the person telling the story. When coding, prioritize a perspective that is NOT the breast cancer organization.] 1. **Breast cancer survivor** 2. **Breast cancer survivor’s family or friends** 3. **Mixed (i.e., survivor + family or friends)** 4. **Journalists/news media** 5. **Breast cancer organization**

This CODEBOOK has been applied consistently across the examples, and it covers all the necessary narrative elements and perspectives without redundancy. Therefore, no changes are needed.

Agent 2: Based on the coding and discussion results, the original CODEBOOK appears to adequately fit the current examples. Each rule in the CODEBOOK has been applied at least once, and the rules are unique with minimal overlap. Therefore, there is no need to revise the CODEBOOK. Here is the original CODEBOOK:

CODEBOOK:

Codebook for Coders (July 1, 2021)

Narrative Elements in Facebook Posts by Breast Cancer Organizations

- **Narrative Event(s) related to breast cancer (plot)**

- [Code the occurrences of all events in the post, such as “3”, “2,4”, or “2,4,5”]

1. Prevention
2. Detection and/or diagnosis
3. Treatment
 - Receiving treatment (e.g., getting the IV chemo, lying in the hospital bed)
 - Treatment effects (e.g., bald head, flat chest, wearing a head wrap)
 - Treatment milestone or completion (e.g., ringing the chemo bell, showing radiation therapy completion certificate).
4. Survivorship - includes complete remission/cancer free; recurrence; a second cancer; and death.
5. Fundraising, any prosocial or philanthropic activities

- **Narrator perspective (narrator_perspective) (evaluated based on the original post; choose one)**

- [Narrator is the person telling the story. When coding, prioritize a perspective that is NOT the breast cancer organization.]

1. Breast cancer survivor
2. Breast cancer survivor’s family or friends
3. Mixed (i.e., survivor + family or friends)
4. Journalists/news media
5. Breast cancer organization

This version of the CODEBOOK covers all cases and patterns in the examples provided, and each rule has been applied appropriately. Therefore, no changes are necessary.

Mediator: ### Summary of Opinions

Both social scientists agree that the original CODEBOOK is adequate and does not require any revisions. They both emphasize that:

1. The original CODEBOOK has been applied consistently across the examples.
2. Each rule in the CODEBOOK has been applied at least once.
3. The rules are unique with minimal overlap.
4. The CODEBOOK covers all necessary narrative elements and perspectives without redundancy.

Updated CODEBOOK

Based on the consensus that no changes are needed, here is the original CODEBOOK for reference:

CODEBOOK:

Codebook for Coders (July 1, 2021)

Narrative Elements in Facebook Posts by Breast Cancer Organizations

Narrative Event(s) related to breast cancer (plot) [Code the occurrences of all events in the post, such as “3”, “2,4”, or “2,4,5”] 1. **Prevention** 2. **Detection and/or diagnosis** 3. **Treatment** - Receiving treatment (e.g., getting the IV chemo, lying in the hospital bed) - Treatment effects (e.g., bald head, flat chest, wearing a head wrap) - Treatment milestone or completion (e.g., ringing the chemo bell, showing radiation therapy completion certificate). 4. **Survivorship** - includes complete remission/cancer free; recurrence; a second cancer; and death. 5. **Fundraising, any prosocial or philanthropic activities**

Narrator perspective (narrator_perspective) (evaluated based on the original post; choose one) [Narrator is the person telling the story. When coding, prioritize a perspective that is NOT the breast cancer organization.] 1. **Breast cancer survivor** 2. **Breast cancer survivor’s family or friends** 3. **Mixed (i.e., survivor + family or friends)** 4. **Journalists/news media** 5. **Breast cancer organization**

Confirmation

Dear Social Scientists,

Based on your feedback, the consensus is that the original CODEBOOK is adequate and does not require any revisions. Here is the original CODEBOOK for your reference.

Please confirm if you agree with this conclusion and the provided CODEBOOK. If there are any further suggestions or concerns, kindly let me know.

Thank you for your collaboration.

Best regards, [Your Name]

Agent 1: ### Confirmation

Dear Social Scientists,

Based on your feedback, the consensus is that the original CODEBOOK is adequate and does not require any revisions. Here is the original CODEBOOK for your reference:

CODEBOOK:

Codebook for Coders (July 1, 2021)

Narrative Elements in Facebook Posts by Breast Cancer Organizations

Narrative Event(s) related to breast cancer (plot) [Code the occurrences of all events in the post, such as “3”, “2,4”, or “2,4,5”] 1. **Prevention** 2. **Detection and/or diagnosis** 3. **Treatment** - Receiving treatment (e.g., getting the IV chemo, lying in the hospital bed) - Treatment effects (e.g., bald head, flat chest, wearing a head wrap) - Treatment milestone or completion (e.g., ringing the chemo bell, showing radiation therapy completion certificate). 4. **Survivorship** - includes complete remission/cancer free; recurrence; a second cancer; and death. 5. **Fundraising, any prosocial or philanthropic activities**

Narrator perspective (narrator_perspective) (evaluated based on the original post; choose one) [Narrator is the person telling the story. When coding, prioritize a perspective that is NOT the breast cancer organization.] 1. **Breast cancer survivor** 2. **Breast cancer survivor's family or friends** 3. **Mixed (i.e., survivor + family or friends)** 4. **Journalists/news media** 5. **Breast cancer organization**

Please confirm if you agree with this conclusion and the provided CODEBOOK. If there are any further suggestions or concerns, kindly let me know.

Thank you for your collaboration.

Best regards, [Your Name]

Agent 2: ### Confirmation

Dear Social Scientists,

Thank you for your thorough analysis and feedback. Based on our discussions and the consistent application of the original CODEBOOK across various examples, we have concluded that the original CODEBOOK is adequate and does not require any revisions. Here is the original CODEBOOK for your reference:

CODEBOOK:

Codebook for Coders (July 1, 2021)

Narrative Elements in Facebook Posts by Breast Cancer Organizations

Narrative Event(s) related to breast cancer (plot) [Code the occurrences of all events in the post, such as “3”, “2,4”, or “2,4,5”] 1. **Prevention** 2. **Detection and/or diagnosis** 3. **Treatment** - Receiving treatment (e.g., getting the IV chemo, lying in the hospital bed) - Treatment effects (e.g., bald head, flat chest, wearing a head wrap) - Treatment milestone or completion (e.g., ringing the chemo bell, showing radiation therapy completion certificate). 4. **Survivorship** - includes complete remission/cancer free; recurrence; a second cancer; and death. 5. **Fundraising, any prosocial or philanthropic activities**

Narrator perspective (narrator_perspective) (evaluated based on the original post; choose one) [Narrator is the person telling the story. When coding, prioritize a perspective that is NOT the breast cancer organization.] 1. **Breast cancer survivor** 2. **Breast cancer survivor's family or friends** 3. **Mixed (i.e., survivor + family or friends)** 4. **Journalists/news media** 5. **Breast cancer organization**

Please confirm if you agree with this conclusion and the provided CODEBOOK. If there are any further suggestions or concerns, kindly let me know.

Thank you for your collaboration.

Best regards, [Your Name]

Judge: Same Agreement: True

F Use of Generative AI

To enhance clarity and readability, we utilized OpenAI o1 exclusively as a language polishing tool. Its role was confined to proofreading, grammatical correction, and stylistic refinement—functions analogous to those provided by traditional grammar checkers and dictionaries. This tool did not contribute to the generation of new scientific content or ideas, and its usage is consistent with standard practices for manuscript preparation.