# Sentimental Agents: Exploring Deliberation, Cognitive Biases, and Decision-making in LLM-based Multiagent Systems

Elizabeth A. Ondula[2,*,†], Daniele Orner[1,†], Nick Mumero Mwangi[1] and Casandra Rusti[2]

[1]*Brave Venture Labs*
[2]*University of Southern California, Los Angeles, USA*

**Abstract**

How does sentiment affect deliberative opinion dynamics in multi-agent systems using Large Language Models (LLMs)? In this paper, we introduce *Sentimental Agents*, a framework designed to study collaborative decision-making in a society of agents, each equipped with a distinct Mental Model of Self. We propose a method to integrate sentiment analysis and a non-Bayesian update mechanism, to analyze and interpret agents' beliefs and interactions systematically. This method allows us to observe the volatility of the sentiment associated with different agent statements, as well as the change in opinion throughout the agents' conversation. We further use it to model and compare collaborative decision-making approaches. We situate these agents in a simulated Human Resource recruiting environment as a case study to evaluate a candidate's fit for a role. We present a set of metrics to assess the quality of the agents' output. Finally, we explore cognitive biases in the agents' individual and collective opinion formation, a fundamental step to enhance decision-making capabilities and mitigate distortions in the system and the agents' collective reasoning.

**Keywords**

Multi-Agent Systems, Large Language Models, Sentiment Analysis, Cognitive Biases, Decision-Making, Opinion Dynamics,

## 1. Introduction

Multi-agent systems (MAS), composed of interactive agents have been pivotal in modeling social phenomena, decision-making processes and collaborative tasks. Large Language Models (LLMs) such as GPT-4 [1] have opened new possibilities for exploring complex social dynamics through the simulation of linguistic interactions among agents. These models can provide the necessary capabilities for simulating communication scenarios. Integrating LLMs into MAS facilitates the study of conversations and interaction patterns in a more detailed manner.

LLMs have demonstrated exceptional performance in generating text that embodies sentiment and in executing sentiment analysis tasks [2]. However, the effect of sentiment on deliberative opinion dynamics within an artificial society of agents is a domain that has not yet been fully explored. Traditional agent models may not adequately account for the influence of behavioral states like sentiment and cognitive biases on the decision-making process. Our work adopts a nuanced approach to understanding how the output of LLM agents influences one another within these frameworks.

We introduce *Sentimental Agents*, a framework designed to study and analyze collaborative decision processes. These agents are not only equipped with language capabilities but also possess a unique *Mental Model of Self*. This allows them to process and exhibit behaviors that can offer a comprehensive view of how opinions are formed and evolve in a multi-agent setting.

Our system is designed primarily to observe and describe agents' behavior, rather than to design or direct it. We do not currently include objectives, reward functions, utility metrics or payoffs in our model. The focus is on the natural evolution of interactions among agents without imposing external incentives or goals. Our study concentrates on non-strategic interactions. Unlike strategic agents, which model the behavior of others and act based on these predictions, our non-strategic agents do not possess such models. This distinction is crucial as it means our agents are not engaging in behaviors such as scheming or deceiving to achieve a specific objective. If LLM-based multi-agent systems are ultimately to be used to support decision-making, it is critical to understand and explain how their decisions are made. This is especially true in the hypothetical case of such systems being designed to evaluate, rank or recommend humans. At present, there are no unified solutions that can system-

*Fourth Workshop on Knowldege-infused Learning, August 25, 2024, Barcelona, Spain*
*Corresponding author.
†These authors contributed equally.
ondula@usc.edu (E. A. Ondula); daniele@braveventurelabs.com (D. Orner); nick@braveventurelabs.com (N. M. Mwangi); rusti@usc.edu (C. Rusti)
https://eondula.github.io/ (E. A. Ondula); https://bravelabs.ai/ (D. Orner); https://bravelabs.ai/ (N. M. Mwangi); https://www.linkedin.com/in/casandrarusti/ (C. Rusti)
0000-0003-0403-0306 (E. A. Ondula); 0009-0005-1264-1985 (D. Orner); 0009-0004-6654-2635 (N. M. Mwangi); 0009-0007-5668-1991 (C. Rusti)

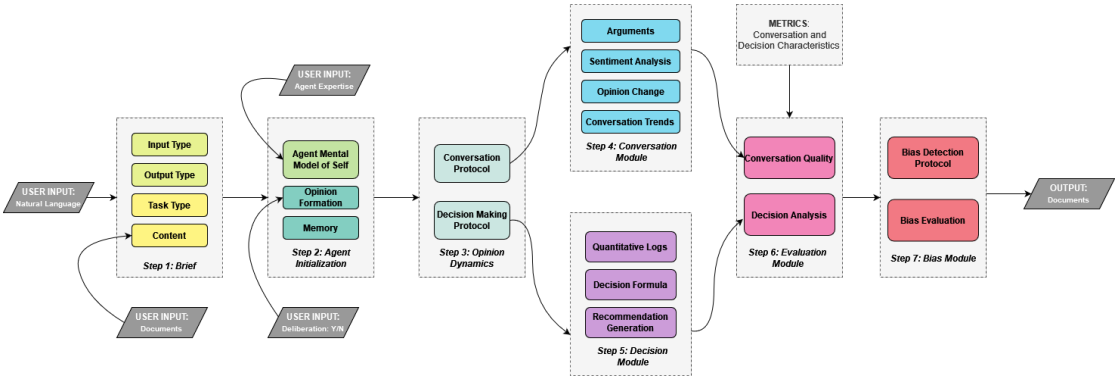



**Figure 1:** *The Sentimental Agents framework consists of 7 modules: The Brief, Agent Initialization, Opinion Dynamics, Conversation, Decision, Cognitive Bias and Evaluation Modules.*

atically analyze the opinions and interactions of these agents, and the potential correlation between the two. To remediate this, we make the following key contributions:

- We develop a framework, *Sentimental Agents* [], to explore and study collective decision-making processes in a society of agents.
- We propose using sentiment analysis as a method to quantify content generated by LLM-based agents for evaluation and recommendation tasks.
- We propose a method to apply a non-Bayesian model for opinion dynamics within a multi-agent system. This offers a perspective on how opinions are formed and altered in a sentiment-driven environment.

Additionally, this work introduces metrics for assessing the quality of conversation and decision-making in language model-based multi-agent systems. These metrics, namely nuance, platitudinal score, drift, and defensibility, offer a toolkit for evaluating the effectiveness of such systems in diverse scenarios. Furthermore, we evaluate cognitive biases including negativity, positivity, and saliency biases. This assessment offers valuable insights into the cognitive influences and tendencies within multi-agent decision-making processes. Finally, the framework is applied in a simulated Human Resource recruiting environment, serving as a practical case study. This application not only validates the theoretical model but also highlights the practical potential of the approach in real-world settings.

# 2. Related Works

## 2.1. Multi-Agent Collaboration

In the study of multi-agent systems, understanding how agents collaborate to achieve collective objectives is essential. One interesting approach, explored [3] examines the use of LLMs in multi-agent settings with a focus on "Theory of Mind" (TOM), which is the ability of an agent to understand and predict the mental states and intentions of others. Although crucial for collaboration, our focus looks more at how agents make decisions rather than understanding others' mental states. Other studies, like those of [4] look at how agents can debate and make collective decisions using a method known as gradual semantics, where agents exchange arguments and progressively update their opinions to reach a shared decision. Our approach is different in that it explains the agent interactions and decision processes leveraging a mental model of self and sentiment tracking. Further, our agents don't have access to other agents' memories. [5] explores how agents coordinate in complex tasks that necessitate both working together on the same task (cooperation) and dividing the task into smaller parts to be done individually (divide-and-conquer). This study highlights the need for flexible strategies to manage tasks that require both joint and individual efforts, differing from our work which doesn't focus on specific task coordination but rather on general deliberations on various topics. Similarly, [6] demonstrate the potential of collaborative mechanisms with LLMs in enhancing social interactions among agents, providing valuable insights into how these technologies can foster collaborative intelligence within multi-agent settings.

## 2.2. LLM-based Multi-Agent Frameworks

An LLM-based agent is defined as an AI system comprising three core components: the brain, perception, and action modules [7]. The brain module stores knowledge and memories, facilitating information processing and decision-making, essential for reasoning and handling new tasks. The perception module extends the agent's sensory capabilities to include textual, auditory, and vi-

sual modalities. This enhances its understanding of the environment. The perception module extends the agent's sensory capabilities to include textual, auditory, and visual modalities. This enhances its understanding of the environment. The action module enables the agent to perform physical tasks and interact with its environment. In terms of operating mechanism, the agent use natural language for communication, with the brain processing information from the perception module to form strategies and make decisions. In our work, we introduce the concept of a *Mental Model of Self (MMS)*. This concept has been discussed in social psychology [8]. It refers to an integrated theory and understanding that an agent forms to organize and make sense of one's self-knowledge, experiences and memories into broader principles that can guide anticipation of future behaviors and consequences. In our implementation, it serves as an important organizational function in making sense of self-knowledge.We summarize and show differences between the *Sentimental Agents* framework and prior works in Table 1.

### 2.3. Non-strategic Multi-Agent Systems

**Opinion dynamics** has been extensively explored for over six decades, predominantly in the fields of sociology and psychology. It delves into the mechanisms and principles that dictate the formation and alteration of individual opinions under the influence of others. This involves examining a range of models and frameworks to comprehend collective behaviors and the process of consensus formation [9]. Our work focuses on a non-strategic model within opinion dynamics, meaning the model does not incorporate game theory principles, nor does it involve agents optimizing specific utilities.

**Non-Bayesian updating**, in this context, signifies a process wherein opinions are modified not based on a factual or probabilistic framework that converts prior probabilities into posterior probabilities. Instead, this approach entails agents updating their opinions influenced by the views of others, without basing these on an unknown state of nature. The updating mechanism in such models can be either synchronous, where all agents update their opinions simultaneously, or asynchronous, where updates occur at different times. A recent survey categorizes and discusses various models prevalent in existing literature [10].

We further use **Sentiment Analysis** to investigate opinions which manifest as either positive or negative [11]. Studies have shown that generative models, such as Large Language Models (LLMs), are capable of producing text, which can include opinions with specific sentiments, depending on their application [12].

### 2.4. Evaluating LLM-based Systems

Evaluation for LLMs is emerging as a discipline to assess the performance of different of AI systems. Currently, for LLMs, there is no single benchmark or protocol that emerges as universally superior. This reflects the diversity of tasks and model capabilities. [2] provides an exhaustive summary and discussion based on existing works. This work covers evaluation tasks, methods and benchmarks that are crucial for assessing the performance of LLMs. In our work, we adopt a nuanced approach to evaluation. We define specific metrics to assess the conversation quality. These metrics include *nuance, platitudinal score, drift* and *defensibility* scores, which are detailed in Section 5.6.

## 3. Preliminaries

### 3.1. Conversation protocols

Consider a conversational simulation system with a set of agents denoted as $\mathscr{M} = \{m_1, m_2, \ldots, m_n\}$. Each agent $m_i \in \mathscr{M}$ is initialized with a Mental Model of Self (MMS) and a memory component for storing an opinion log. In this system, the engagement among agents in each round $t$ is ordered with equal participation.

**Definition 1.** ***Argument (***$A$***)*** *is a component of an opinion that contributes to its overall sentiment.*

For each argument $A$ a sentiment value $S_A$ is assigned, mapping the argument to a spectrum of sentiment values (positive, negative, neutral, and their intensities):

$$S_A = f(A) \tag{1}$$

where $f : A \mapsto S_A$ is the sentiment mapping function for arguments.

**Definition 2.** *$O(m_i, t)$ is the opinion $m_i$ in a given round $t$ is a set of arguments $A$.*

The sentiment of an opinion $S_O$ is the average of the sentiment values $S_A$ of all its arguments:

$$S_O(m_i, t) = \frac{1}{|O(m_i, t)|} \sum_{A \in O(m_i, t)} S_A \tag{2}$$

The **Ordered Engagement** in the system is represented by a function $E : \mathscr{M} \times t \to m_i$, which establishes the speaking order of agents in each round $t$. Under this model, each agent $m_i$ contributes exactly one opinion per round. The collective state of opinions at any given round $t$ is represented as a vector:

$$X_E(t) = [O(m_1, t), O(m_2, t), \ldots, O(m_n, t)] \tag{3}$$

In each conversation round $t$, the sentiment value $S_O m_i, t$ for each agent $m_i$ is updated to reflect the sentiment of the

| Related Work | Sentiment Analysis | Engagement type | Memory | Decision module | Bias Evaluation |
|---|---|---|---|---|---|
| [13] | No | Ordered | Belief | No | Confirmation bias |
| [14] | No | Ordered | Store/Retrieve | Yes | No |
| [15] | No | Ordered | Internal critic | Yes | Fact-checking |
| [16] | No | Varies | Chat history | Yes | No |
| [17] | No | Ordered | Specialized roles | Yes | No |
| [18] | No | Ordered | User-driven | Yes | User-preference |
| [19] | No | Ordered | Rationale analysis | Yes | Credibility check |
| [20] | No | Ordered | Dynamic Memory | No | Opinion classifier |
| **Sentimental Agents** | Yes | Ordered | Opinion logs | Yes | Cognitive bias |

**Table 1**
*A comparison of different language model-based multi-agent frameworks.*

newly formed opinion. This process considers the sentiment values $S_A$ of the arguments within the opinion $O$. The sentiment update is executed using a Non-Bayesian method, mathematically represented by:

$$S_{i_O}(t) = \alpha \cdot \left( \frac{1}{|A|} \sum_{A \in O(m_i,t)} S_A \right) + (1-\alpha) \cdot S_{i_O}(t-1) \quad (4)$$

Here, $S_{i_O}(t)$ represents the average sentiment of all the arguments expressed by agent $m_i$ at round $t$, with each argument $A$ having its sentiment value $S_A = f(A)$. The parameter $\alpha$ is a weighting factor that determines the influence of the new opinion's average sentiment on the agent's updated sentiment.

The change in sentiment $\Delta S_{i_O}(t)$ for agent $m_i$ is then calculated as the absolute difference between the updated sentiment value $S_{i_O}(t)$ at round $t$ and the agent's previous sentiment value $S_{i_O}(t-1)$ at round $t-1$:

$$\Delta S_{i_O}(t) = |S_{i_O}(t) - S_{i_O}(t-1)| \quad (5)$$

## 3.2. Collective decision protocols

When the conversation ends we take the total sentiment. We have the final sentiment score and we have the average of the $S_0$ for the gut feeling protocol

**Definition 3.** ***Borda Count Protocol****: A method to collectively rank a list of items, given each individual's order of preference.*

Given $n$, the number of items, each agent $m_i$ ranks these items. The point assignment for an item $j$ by agent $m_i$ is $P_{m_i,j}$, with the top-ranked item receiving $n$ points and the last receiving 1 point. The total points for each item $j$ is calculated as $T_j = \sum_{i=1}^{|\mathcal{M}|} P_{m_i,j}$, and items are ranked in descending order of their total points $T_j$.

**Definition 4.** ***Tiered List Protocol****: A method to collectively classify a list of items in 3 tiers, given the items that each individual can't accept, and the items they like the most.*

The Valence $V_j$ for each item $j$ is determined based on the sentiment of opinion $S_O$. For $S_O < -0.5$, $V_j = -1$; for $-0.5 \leq S_O \leq 0.5$, $V_j = 0$; and for $S_O > 0.5$, $V_j = 1$. Items are classified into three tiers according to $V_j$: Tier 1 for $V_j = 1$, Tier 2 for $V_j = 0$, and Tier 3 for $V_j = -1$.

**Definition 5.** ***Gut-feeling List Protocol****: A method to collectively rank a list of items based on the confidence of individuals' feeling toward each item.*

The volatility $v_{m_i,j}$ of agent $m_i$'s sentiment towards item $j$ over several rounds is calculated. Conviction $I_{m_i,j}$ is derived as a function of both volatility $v_{m_i,j}$ and the final sentiment score $S_{m_i,j}$ for item $j$. The Gut-feeling list is then generated using a Borda count based on $I_{m_i,j}$ for each item across all agents, and items are ranked based on the total Conviction points $T_j^I = \sum_{i=1}^{|\mathcal{M}|} I_{m_i,j}$ in descending order.

# 4. Applying the Framework

Our framework is applied to a simulated environment inspired by Human Resource recruiting to evaluate the effectiveness of *Sentimental Agents*. These agents are designed to generate opinions reflecting their unique expertise, contributing to collective decision-making. The simulation explores opinion formation and decision-making processes within an LLM-based multi-agent setting, mirroring real-world HR recruitment where employers assess candidates through discussions with various experts. In this context, LLM-based agents are expected to engage in conversation and form diverse opinions that influence their decision-making in a simulated recruiting scenario.

## 4.1. Configuration

In the HR recruiting simulation, advisor agents analyze candidates' CVs and engage in discussions to provide opinions about each candidate. These agents, with expertise in roles like Chief Financial Officer (CFO), Vice President of Engineering, and Recycling Plant Manager,

evaluate profiles and generate text reports. They also score candidates and, through collective decision-making protocols like the Borda Count, rank candidates or select the top performers.

#### 4.1.1. Dataset

We sourced our dataset from the study conducted by [21]. This dataset is a collection of resumes represented in a multi-label format. To facilitate easy access and integration of this dataset into our framework, we have developed a script that automates the process of downloading and parsing the data.

## 5. Sentimental Agents Framework

The system design as shown in Fig 1, consists of 7 modules. We describe each of them here.

### 5.1. Brief Module

The module provides a configuration interface for system initialization with four components: `input type`, `output type`, `task type`, and `context`. It handles single and multiple item formats for input and output and requires user-defined context specifying `task object` and `subject`, with optional Knowledge base integration. Predefined rules in the module automatically associate Input, Output, and Task Types. The logic enforces specific task types *Evaluate, Score, Classify* for single-item inputs and broader tasks for multi-item inputs. For `rank` tasks, the output is structured as a list to match task requirements. This design ensures alignment between input/output formats and system functionality.

### 5.2. Agent Initialization

The Agent Initialization module includes two main elements: Mental Model of Self (MMS) and Memory with qualitative and quantitative opinion logs. It configures agent interaction types for opinion formation as dynamic or independent. The module requires user input to set the number of agents and their expertise, which informs the creation of detailed agent profiles, including priorities, objectives, and evaluation criteria. Figure 3 shows an instance of an MMS. Key parameters include the *tolerance level*, affecting opinion change propensity, and the *drift* metric, which tracks MMS variability. Strategies for maintaining agent consistency involve controlled character prompts and setting MMS prompt *temperature*. The opinion formation, generated via boolean input, influences the nature of agents' decision-making processes. Each agent's opinion log is stored in a central memory system, ensuring decision-making is based on comprehensive and transparent data. Figure 2 shows agent initialization prompt.

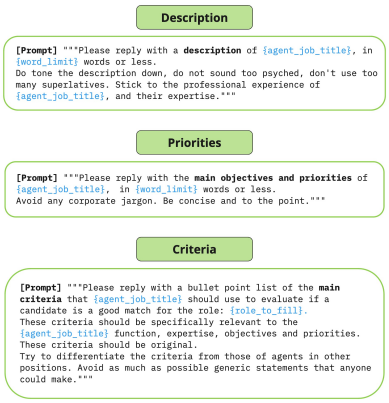



**Figure 2:** *Series of prompts used to create a group of agents Mental Model of Self, for one instance of the system*

### 5.3. Opinion Dynamics Module

This module coordinates agent conversations and decision-making, consisting of conversation and decision-making protocols. It focuses on: defining the number of agents, engagement type and stopping mechanism. The current implementation employs ordered engagement with equal participation. For the stop mechanism, the module uses a non-strategic approach, differentiating from strategic interactions. In this non-strategic context, conversations conclude based on non-Bayesian updating as shown in Algorithm 1 , where they end once agents' opinions reach stability. This contrasts with strategic interactions, which involve different mechanisms like rewards or objectives.

### 5.4. Conversation Module

This module analyzes agents' statements in conversations, comprising four components: Argumentation, which breaks down statements into arguments for qualitative logging; Sentiment Analysis, which evaluates and quantitatively logs the sentiment of each argument; Opinion Change, using non-Bayesian updating to monitor sentiment shifts; and Conversation Trends, gauging significant changes across rounds to infer opinion stabilization and conversation conclusion.

### 5.5. Decision Module

The Decision Making Protocol module is designed to accommodate various decision-making protocols, including Borda Count, Tiered List, and Gut Feeling List, as detailed

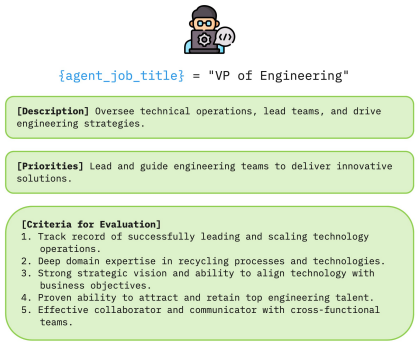



**Figure 3:** *An instance of an agent's Mental Model of Self in a simulated HR environment. In this case, the agent took a Job Title as input, and generated a Description, Priorities, and Evaluation Criteria for a given Job Description.*

in the preliminaries (Section 3). It operates by capturing the final sentiment of each agent and the average sentiment throughout the conversation. The functionality and outcomes of these different decision-making processes are further explored and discussed in the results (Section 6).

### 5.6. Evaluation and Cognitive Bias Modules

This module evaluates the quality of conversations through various metrics.

- **Nuance**: Examines the diversity of themes and perspectives, quantified by the number of topics identified within individual statements or the entire conversation.
- **Platitudinal Score**: Calculated using cosine similarity, it measures the uniqueness of outcomes in the conversation rounds, with higher scores indicating less similarity between different runs.
- **Drift**: Assesses the stability of each agent's Mental Model of Self, monitoring the relevance of results to the advisors' profiles and checking for consistency throughout the conversation.
- **Defensibility**: Evaluates the strength and evidence backing of the agents' arguments, ensuring they are well-supported and referenceable.

In this research, we examine three cognitive biases: negativity, positivity, and saliency. Negativity bias might lead agents to give undue weight to adverse opinions [22] [23], while positivity bias could result in an overemphasis on favorable views [24] [25]. Saliency bias, on the other hand, might cause agents to focus on the most prominent or emotionally striking aspects of an opinion, potentially overshadowing other relevant information [26].

**Algorithm 1** Non-Bayesian Updating

**for** each round $t$ **do**
  **for** each agent $m_i \in \mathscr{M}$ **do**
    **if** $t > 0$ **then**
      Equation 5 $\Delta S_{i_O}(t) = |S_{i_O}(t) - S_{i_O}(t-1)|$
      Equation 4 $S_{i_O}(t) = \alpha \cdot S_O(m_i, t) + (1-\alpha) \cdot S_{i_O}(t-1)$
    **end if**
  **end for**
  **if** all $\Delta S_{i_O}(t) <$ threshold for each $m_i \in \mathscr{M}$ or $t =$ max_rounds **then**
    Set *conversation_active* to False
  **end if**
  Increment $t$
**end for**

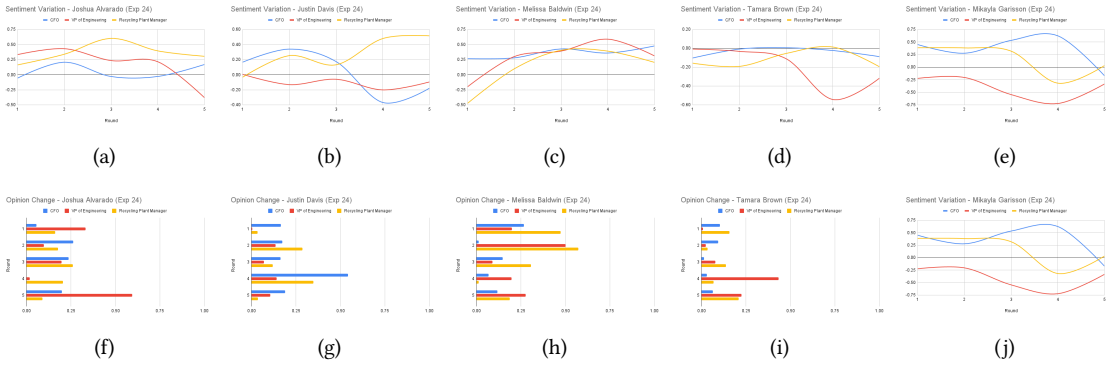
## 6. Experiments

In this study, we aim to investigate the dynamics of sentiment and opinion formation in an LLM-based multi-agent system. We focus on understanding how agents' opinions evolve through deliberation, and how sentiment influences their decision-making processes. Our research questions are as follows:

1. How do agents' opinions change as a result of deliberating with each other, and can we quantify these changes?
2. Do agents adopt each other's arguments during the deliberation process, and can we observe this in qualitative results?
3. Does the sentiment of an argument (valence, arousal) affect its adoption by other agents?
4. Do agents exhibit cognitive biases in their opinion formation, and how can we identify and mitigate these biases?

### 6.1. Experimental Setting

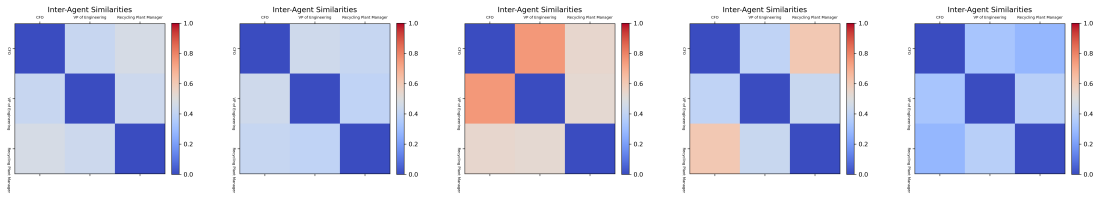
In our experiments, we conducted the simulation with 3 agents and 10 candidates. We used the data set within the simulation environment described in Section 4. For the LLM, we used the gpt-3.5-turbo-0613 version of ChatGPT [27]. For the result shown, the language model parameters were set as alpha $\alpha = 0.5$, tolerance = 0.00001, and temperature = 1.5.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

**Figure 4:** *Sentiment change, and corresponding Opinion Change in conversations for two different candidates. Each conversation stops after five rounds.*



(a) (b) (c) (d) (e)

**Figure 5:** *Platitudinal Score: The uniqueness of outcomes in the conversation rounds among agents (a lower Score indicates a more original contribution)*

## 6.2. Results

### 6.2.1. Evaluation Metrics

The non-Bayesian updating data from the simulation, shown in Figure 4, reveals sentiment fluctuations among agents. For instance, Figure 5a shows the VP of Engineering exhibiting the most dramatic change, especially in the final round. This volatility, captured by sentiment and change metrics, highlights the dynamic nature of opinion formation in multi-agent conversations and suggests that agents' opinions evolve and respond to the unfolding discourse, emphasizing the effectiveness of non-Bayesian updating in capturing real-time perspective shifts. [1]

**Platitudinal score**. The inter-agent similarities heatmap shown in Figure 5 reveals a contrast in sentiment alignment among the agents. This divergence contributes to an overall lower platitudinal score for this specific run for the given candidate. Such diversity in sentiment, as captured by the platitudinal metric, underscores the variation in decision-making approaches within the agent group, emphasizing the balance between consensus and individual thought in the simulation outcomes.

[1]For brevity, we only show results for 5 candidates, but the experiment was conducted with 10 candidates for the platitudinal scores, Sentiment and Opinion change

**Drift scores**. In Table 3 it is observed that the CFO agent generally exhibits moderate drift, while the VP of Engineering (VPE) and the Recycling Plant Manager (RPM) show higher drift values, suggesting a more dynamic adaptation of their MMS in response to the conversation. This variability in drift signifies the agents' differing levels of adaptability and potential reevaluation of their initial stances

| Candidate | CFO | VPE | RPM |
|---|---|---|---|
| Kimberly Carr | 0.4504 | 0.7556 | 0.7138 |
| Melissa Morgan | 0.6224 | 0.6308 | 0.5810 |
| Mikayla Garrison | 0.3254 | 0.5878 | 0.5720 |
| Emily Marshall | 0.4678 | 0.7998 | 0.7390 |
| Justin Davis | 0.3458 | 0.3638 | 0.3940 |
| Tamara Brown | 0.3842 | 0.6030 | 0.6574 |
| Taylor Mahoney | 0.3814 | 0.4794 | 0.4154 |
| Joshua Alvarado | 0.3756 | 0.5238 | 0.5788 |
| Melissa Baldwin | 0.4228 | 0.7988 | 0.5714 |
| James Wallace | 0.4240 | 0.6342 | 0.6926 |

**Table 2**
*Agent Drift Values for hypothetical candidates*

**Nuance Scores** We use Latent Dirichlet Allocation

(LDA) to extract topics from text statements. The data is preprocessed by tokenization and removal of stop words and unwanted words. A dictionary and corpus are constructed using the Gensim library. The LDA model identifies 5 topics, with the top 10 words per topic being most significant. Figure 6 and **??** show the number of unique words per topic and word clouds for each candidate, respectively.

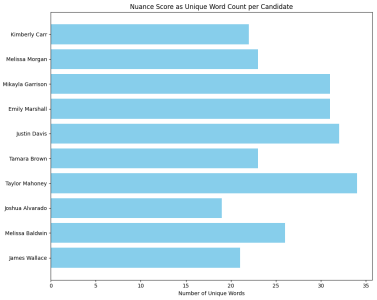



**Figure 6:** *The nuance score for each candidate by showing the number of unique words used across all topics.*

**Defensibility Scores** Candidate resumes are processed through the Langchain embedding[2] and transformed into a format suitable for detailed analysis. The llama index libraries, `VectorStoreIndex` and `ServiceContext`, are used to create an indexed repository of the vectorized documents. This index serves as a searchable database, allowing efficient retrieval of text segments that are contextually similar to a given input. When evaluating agents' arguments, the indexed space is searched to find text segments from the resumes that closely match the argument. The similarity between an agent's argument and the retrieved text is quantified as a score, with higher scores indicating stronger support for the argument. If no relevant text is found, a score of zero is assigned, suggesting an unsupported argument.

### 6.2.2. Cognitive Bias Testing

We hypothesize that agents' updates in sentiment during conversational rounds might be influenced by their peers' positive, negative, or prominent opinions. To investigate this, we chart each agent's sentiment change from the second round onwards, against the recent sentiments of other agents. This analysis reveals the correlation between an agent's changing sentiment and the influence of peer opinions.

We apply Ordinary Least Squares (OLS) regression to analyze nega tive and positive sentiments separately, setting the y-intercept at zero to indicate that neutral peer statements might not impact an agent's sentiment. Analyzing the regression's strength ($R^2$) and the slope, as well as the data point distribution, provides insights into the cognitive tendencies of the agents. Additionally, by adjusting our three parameters, *alpha*, *tolerance*, and *temperature*, we aim to better understand how these factors affect agents' cognitive biases. This study offers important insights into the decision-making processes in multi-agent systems, particularly in sentiment-influenced contexts.

In our sensitivity analysis, we varied key parameters: setting *alpha* to 0.3, 0.5, and 0.7; *tolerance* to 0.001, 0.005, and 0.0001; and *temperature* to 0.7, 1, and 1.5, to evaluate their impact on sentiment changes. The outcome, depicted in Figure 7 for ten random candidates, provides insight into negativity and positivity biases through the slopes of the OLS regressions. Our findings on this variation of model parameters show a modest positivity bias, evidenced by the positive slope being approximately 29% steeper than its negative counterpart. A slight positivity or negativity bias trend persisted across varied parameter settings, with some scenarios, notably *alpha* = 0.3, *tolerance* = 0.005, and *temperature* = 1.5, showing a more pronounced positivity bias with a slope more than twice as steep on the positive side than on the negative side.

The absence of saliency bias was noted in all experiments, as indicated by slopes remaining below 1. Linear regression was determined as the most suitable model based on our evaluation of the $R^2$ values. Notably in the shown experiment, agents displayed a tendency towards expressing stronger negative sentiments, with the most negative reaching -0.76, compared to a maximum positive sentiment of 0.62. This inclination towards stronger negative expressions was marked in most scenarios. Additionally, the *alpha* parameter was observed to significantly influence sentiment ranges, with lower *alpha* values yielding more constrained ranges.

For future studies, we aim to extend our examination of the cognitive bias to larger candidate sample sizes. This expansion will enable us to deepen our understanding of how parameter tuning influences cognitive biases and decision-making processes within our framework.

### 6.2.3. Collective decision-making

The decision-making data reveals diverse agent preferences, as evidenced by the variation in candidate ranks across Borda Count, Tier, and Conviction. We use the average sentiment score, $S_{i_O}(t)$, from equation 4, where $t$ is the last round, as the basis for collective decision-making. While some candidates consistently rank higher or lower, suggesting a consensus on their suitability, discrepancies in ranks among agents could reflect unique valuations of candidate qualities.

Table 4 shows the overall sentiment scores. The CFO shows the highest sentiment score, of 0.46 towards Melissa Baldwin, indicating a strong positive inclination.

[2]https://github.com/langchain-ai/langchain

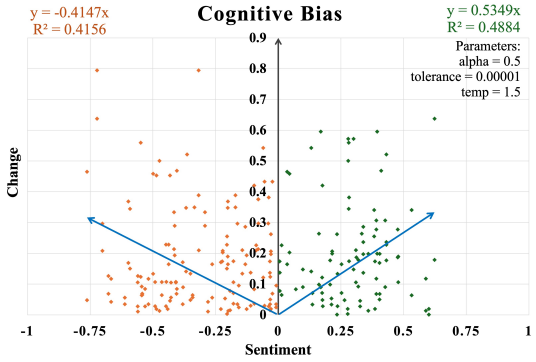



**Figure 7:** *The results of the cognitive bias testing using a sample of ten candidates and model parameters of alpha = 0.5, tolerance = 0.00001, and temperature = 1.5.*

| Candidate | Borda Count | Tiered | Gut-Feeling |
|---|---|---|---|
| Kimberly | 9 | 3 | 10 |
| Melissa | 10 | 3 | 8 |
| Mikayla | 2 | 2 | 1 |
| Emily | 3 | 2 | 3 |
| Justin | 5 | 2 | 5 |
| Tamara | 6 | 2 | 6 |
| Taylor | 8 | 3 | 7 |
| Joshua | 4 | 2 | 4 |
| Melissa | 1 | 2 | 2 |
| James | 7 | 2 | 9 |

**Table 3**
*Candidate Ranking Metrics*

| Final Sentiment Scores | | | |
|---|---|---|---|
| Candidate | CFO | VPE | RPM |
| Kimberly Carr | -0.51 | -0.13 | 0.26 |
| Melissa Morgan | -0.74 | -0.21 | -0.46 |
| Mikayla Garrison | 0.69 | 0.38 | 0.59 |
| Emily Marshall | 0.29 | -0.29 | 0.05 |
| Justin Davis | -0.01 | -0.12 | 0.37 |
| Tamara Brown | 0.11 | 0.36 | 0.20 |
| Taylor Mahoney | -0.49 | -0.55 | -0.41 |
| Joshua Alvarado | 0.28 | 0.58 | 0.25 |
| Melissa Baldwin | 0.46 | 0.34 | 0.17 |
| James Wallace | -0.30 | 0.11 | -0.30 |

**Table 4**
*Final Sentiment Scores for each candidate*

| Candidates Rank | | | | |
|---|---|---|---|---|
| Candidate | CFO | VPE | RPM | Borda Count Rank |
| Kimberly Carr | 9 | 5 | 3 | 4 |
| Melissa Morgan | 10 | 6 | 10 | 2 |
| Mikayla Garrison | 2 | 9 | 4 | 7 |
| Emily Marshall | 3 | 8 | 6 | 4 |
| Justin Davis | 5 | 4 | 2 | 8 |
| Tamara Brown | 6 | 7 | 7 | 3 |
| Taylor Mahoney | 8 | 10 | 9 | 1 |
| Joshua Alvarado | 4 | 3 | 1 | 9 |
| Melissa Baldwin | 1 | 1 | 5 | 10 |
| James Wallace | 7 | 2 | 8 | 4 |

**Table 5**
*Rank of each candidate, including the final Rank taking into account each agent's individual rankings (calculated through Borda count)*

In contrast, the CFO's lowest sentiment score is -0.74 towards Melissa Morgan, signaling a significant negative view. Similarly, the VPE aligns with the CFO in favoring Melissa Baldwin with the highest score of 0.34, but diverges in its lowest sentiment, which is directed towards Taylor Mahoney with a score of -0.55. The RPM, on the other hand, exhibits the most positive sentiment towards Mikayla Garrison with a score of 0.59, while sharing the CFO's negative sentiment towards Melissa Morgan, albeit at a less intense level of -0.46. The sentiment scores from the conversations directly influence the ranking of candidates as shown in Table 5. Applying the Borda Count method to the combined rankings yields a collective decision. Although individual agents might rank candidates differently based on their interactions, the aggregated results provide a more comprehensive assessment. This approach demonstrates how sentiment analysis combined with a voting system could inform hiring decisions in a multi-agent setting.

The intensity of an agent's final sentiment score determines the valence score. In Table 6, this occurred only three times: the CFO attributed a negative valence to two candidates, while the VPE attributed a negative valence to one candidate. Consequently, no candidate was classified as Tier 1, with most classified as Tier 2, except for the three candidates with negative valence, who were classified as Tier 3.

| Valence | | | |
|---|---|---|---|
| Candidate | CFO | VPE | RPM |
| Kimberly Carr | -1 | 0 | 0 |
| Melissa Morgan | -1 | 0 | 0 |
| Mikayla Garrison | 0 | 0 | 0 |
| Emily Marshall | 0 | 0 | 0 |
| Justin Davis | 0 | 0 | 0 |
| Tamara Brown | 0 | 0 | 0 |
| Taylor Mahoney | 0 | -1 | 0 |
| Joshua Alvarado | 0 | 0 | 0 |
| Melissa Baldwin | 0 | 0 | 0 |
| James Wallace | 0 | 0 | 0 |

**Table 6**
*Valence for each candidate*

The sentiment volatility of the agents, as shown in Table 7, was mostly moderate, indicating strong conviction in their opinions. However, there were instances of high

volatility, such as the CFO's sentiment towards Kimberly Carr and Mikayla Garrison, and the VPE's sentiment towards Kimberly, Joshua, and James. The RPM's sentiment was volatile towards Mikayla and Melissa Baldwin. The agents' conviction in their opinions is calculated by dividing the final sentiment by the volatility, with higher values indicating stronger intuition about a candidate's suitability for the role.

| Sentiment Volatility | | | |
|---|---|---|---|
| Candidate | CFO | VPE | RPM |
| Kimberly Carr | 0.58 | 0.64 | 0.40 |
| Melissa Morgan | 0.16 | 0.32 | 0.47 |
| Mikayla Garrison | 0.31 | -0.47 | 0.17 |
| Emily Marshall | 0.17 | 0.42 | 0.21 |
| Justin Davis | 0.49 | 0.16 | 0.34 |
| Tamara Brown | -0.06 | -0.27 | -0.16 |
| Taylor Mahoney | 0.13 | 0.26 | 0.17 |
| Joshua Alvarado | 0.09 | 0.09 | 0.42 |
| Melissa Baldwin | 0.14 | 0.46 | 0.61 |
| James Wallace | 0.39 | 0.54 | 0.39 |

**Table 7**
*Sentiment Volatility for each candidate*

| Conviction | | | |
|---|---|---|---|
| Candidate | CFO | VPE | RPM |
| Kimberly Carr | -0.30 | -.0.08 | 0.11 |
| Melissa Morgan | -0.12 | -0.07 | -0.22 |
| Mikayla Garrison | 0.21 | -0.18 | 0.10 |
| Emily Marshall | 0.05 | -0.12 | 0.01 |
| Justin Davis | 0.00 | -0.02 | 0.13 |
| Tamara Brown | -0.01 | -0.10 | -0.03 |
| Taylor Mahoney | -0.06 | -0.14 | -0.07 |
| Joshua Alvarado | 0.02 | 0.05 | 0.10 |
| Melissa Baldwin | 0.06 | 0.16 | 0.10 |
| James Wallace | -0.12 | 0.06 | -0.12 |

**Table 8**
*Conviction for each candidate (the Sentiment Score of given by agents to a candidate, taking into account the Sentiment Volatility during a conversation about this candidate)*

Gut Feeling Rank for each candidate is a revised ranking that takes into account an agent's conviction in its own sentiment. In Table 9, the Gut Feeling of the RPM toward Joshua is still to rank him in the first place. But the CFO revises its ranking of Kimberley, from the 9th place to the 8th place. The more generous ranking can be interpreted as a result of the "acknowledgement" of the RPM agent that it is not sure of its opinion toward Kimberly.

| Gut Feeling | | | | |
|---|---|---|---|---|
| Candidate | CFO | VPE | RPM | Borda Count Rank |
| Kimberly Carr | 2 | 6 | 2 | 5 |
| Melissa Morgan | 8 | 5 | 10 | 2 |
| Mikayla Garrison | 2 | 9 | 4 | 7 |
| Emily Marshall | 3 | 8 | 6 | 6 |
| Justin Davis | 5 | 4 | 1 | 8 |
| Tamara Brown | 6 | 7 | 7 | 3 |
| Taylor Mahoney | 7 | 9 | 8 | 1 |
| Joshua Alvarado | 4 | 3 | 1 | 9 |
| Melissa Baldwin | 2 | 1 | 4 | 10 |
| James Wallace | 9 | 2 | 9 | 3 |

**Table 9**
*Gut feeling for each candidate (Ranking of candidates that combines both the Sentiment Score, and the Conviction an agent has in this sentiment*

## 7. Conclusion

In this paper, we introduce **Sentimental Agents**, LLM-based agents that generate opinions for collective decision-making within conversational settings. Our proposed framework integrates a non-Bayesian updating mechanism to track sentiment volatility and opinion evolution. In a simulated HR recruiting scenario, we assess these agents' decision-making abilities, noting their diverse opinions and preference shifts over multiple rounds. The findings suggest model parameters, such as alpha and tolerance, significantly influence sentiment expression and thus cognitive bias within the system. This research offers a foundation for advanced tool development applicable to domains such as HR recruiting, medical diagnostics, or educational domains.

## References


[1] R. OpenAI, Gpt-4 technical report. arxiv 2303.08774, View in Article 2 (2023) 13.

[2] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, A survey on evaluation of large language models, arXiv preprint arXiv:2307.03109 (2023). URL: https://arxiv.org/abs/2307.03109.

[3] H. Li, Y. Q. Chong, S. Stepputtis, J. Campbell, D. Hughes, M. Lewis, K. Sycara, Theory of mind for multi-agent collaboration via large language models, arXiv preprint arXiv:2310.10701 (2023).

[4] L. D. de Tarlé, E. Bonzon, N. Maudet, Multiagent dynamics of gradual argumentation semantics, in: 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), 2022.

[5] S. A. Wu, R. E. Wang, J. A. Evans, J. B. Tenenbaum, D. C. Parkes, M. Kleiman-Weiner, Too many cooks: Bayesian inference for coordinating multi-agent collaboration, Topics in Cognitive Science 13 (2021) 414–432.

[6] J. Zhang, X. Xu, S. Deng, Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology

View, 2023. URL: http://arxiv.org/abs/2310.02124, arXiv:2310.02124 [cs].
[7] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al., The rise and potential of large language model based agents: A survey, arXiv preprint arXiv:2309.07864 (2023).
[8] D. Hart, S. Fegley, Social imitation and the emergence of a mental model of self. (1994).
[9] A. Sîrbu, V. Loreto, V. D. P. Servedio, F. Tria, Opinion Dynamics: Models, Extensions and External Effects, in: V. Loreto, M. Haklay, A. Hotho, V. D. Servedio, G. Stumme, J. Theunis, F. Tria (Eds.), Participatory Sensing, Opinions and Collective Awareness, Springer International Publishing, Cham, 2017, pp. 363–401. URL: http://link.springer.com/10.1007/978-3-319-25658-0_17. doi:10.1007/978-3-319-25658-0_17, series Title: Understanding Complex Systems.
[10] M. Grabisch, A. Rusinowska, A survey on nonstrategic models of opinion dynamics, Games 11 (2020) 65.
[11] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, et al., A practical guide to sentiment analysis, volume 5, Springer, 2017.
[12] U. Maqsud, Synthetic text generation for sentiment analysis, in: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2015, pp. 156–161.
[13] G. Betz, Natural-language multi-agent simulations of argumentative opinion dynamics, arXiv preprint arXiv:2104.06737 (2021).
[14] Y. Li, Y. Zhang, L. Sun, Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents, arXiv preprint arXiv:2310.06500 (2023).
[15] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, arXiv preprint arXiv:2305.14325 (2023).
[16] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, Z. Liu, Chateval: Towards better llm-based evaluators through multi-agent debate, arXiv preprint arXiv:2308.07201 (2023).
[17] G. Chen, S. Dong, Y. Shu, G. Zhang, J. Sesay, B. F. Karlsson, J. Fu, Y. Shi, Autoagents: A framework for automatic agent generation, arXiv preprint arXiv:2309.17288 (2023).
[18] J. Park, B. Min, X. Ma, J. Kim, ChoiceMates: Supporting Unfamiliar Online Decision-Making with Multi-Agent Conversational Interactions, 2023. URL: http://arxiv.org/abs/2310.01331, arXiv:2310.01331 [cs].
[19] X. Sun, X. Li, S. Zhang, S. Wang, F. Wu, J. Li, T. Zhang, G. Wang, Sentiment Analysis through LLM Negotiations, 2023. URL: http://arxiv.org/abs/2311.01876, arXiv:2311.01876 [cs].
[20] Y.-S. Chuang, A. Goyal, N. Harlalka, S. Suresh, R. Hawkins, S. Yang, D. Shah, J. Hu, T. T. Rogers, Simulating Opinion Dynamics with Networks of LLM-based Agents, 2023. URL: http://arxiv.org/abs/2311.09618, arXiv:2311.09618 [physics].
[21] K. Jiechieu, N. Tsopze, Skills prediction based on multi-label resume classification using cnn with model predictions explanation, Neural Computing and Applications (2020). URL: https://doi.org/10.1007/s00521-020-05302-x. doi:10.1007/s00521-020-05302-x.
[22] T. A. Ito, J. T. Larsen, N. K. Smith, J. T. Cacioppo, Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations., Journal of personality and social psychology 75 (1998) 887.
[23] P. Rozin, E. B. Royzman, Negativity bias, negativity dominance, and contagion, Personality and social psychology review 5 (2001) 296–320.
[24] M. W. Matlin, D. J. Stang, The Pollyanna principle: Selectivity in language, memory, and thought, Schenkman Publishing Company, 1978.
[25] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, et al., Human language reveals a universal positivity bias, Proceedings of the national academy of sciences 112 (2015) 2389–2394.
[26] M. P. Inderbitzin, A. Betella, A. Lanatá, E. P. Scilingo, U. Bernardet, P. F. Verschure, The social perceptual salience effect., Journal of experimental psychology: human perception and performance 39 (2013) 62.
[27] https://platform.openai.com/docs/models/gpt-3-5, 2024. Accessed: 2024-10-9.