Evaluating Large Language Models on Social Signal Sensitivity: An Appraisal Theory Approach

Zhen Wu, Ritam Dutt, Carolyn Penstein Rosé

Language Technologies Institute, Carnegie Mellon University
{zhenwu,rdutt,cprose}@cs.cmu.edu

Abstract

We present a framework to assess the sensitivity of Large Language Models (LLMs) to textually embedded social signals using an Appraisal Theory perspective. We report on an experiment that uses prompts encoding three dimensions of social signals: Affect, Judgment, and Appreciation. In response to the prompt, an LLM generates both an analysis (Insight) and a conversational Response, which are analyzed in terms of sensitivity to the signals. We quantitatively evaluate the output text through topical analysis of the Insight and predicted social intelligence scores of the Response in terms of empathy and emotional polarity. Key findings show that LLMs are more sensitive to positive signals. The personas impact Responses but not the Insight. We discuss how our framework can be extended to a broader set of social signals, personas, and scenarios to evaluate LLM behaviors under various conditions.

1 Introduction

> *"The limits of my language mean the limits of my world."* (Wittgenstein, 1922)

The increasing integration of Large Language Models (LLMs) into social contexts presents a critical challenge: How effectively can they process and respond to social signals embedded in human language? Social signals, as defined in Poggi and Francesca (2010), are communicative or informative signals that convey insights into social actions (e.g., insulting someone), interactions (e.g., showing responsiveness), emotions (e.g., reflecting joy), attitudes (e.g., exhibiting disgust), and relationships (e.g., showing closeness). These social signals are tools in interaction for maintaining or changing relationships that set the stage for effective human-human interactions, which may shape the responses of LLMs when they engage as participants in hybrid settings involving both humans and LLMs.

This paper illustrates a methodology for systematic investigation of the sensitivity of LLMs to social signals in role-playing scenarios. In particular, the research specifically focuses on social signals grounded in Appraisal Theory (Martin and White, 2005) — Affect, Judgment, and Appreciation. These dimensions facilitate a nuanced understanding of how human language expresses emotions, makes ethical judgments, and appreciates the significance of practices respectively. In particular, the research aims to address two main questions:

- **RQ1:** How sensitive are current LLMs to the encoding of social signals in language, both in terms of ability to explain the encoding of social signals in the text and to reply in ways that are responsive to the signal?
- **RQ2:** As a test of generality across contexts, how and to what extent does sensitivity to social signals change as constraints are placed on LLM behavior, such as introducing a persona as a guiding principle for behavior generation?

The research paradigm is displayed in Figure 1. The framework is meant to assess specific capabilities of LLMs, identify limitations, and address challenges in utilizing sociolinguistic theories in such evaluations. Our contributions are as follows:

- We take an exploratory approach to investigate the sensitivity of LLMs to social signals grounded in Appraisal Theory (Martin and White, 2005).
- Our experimental design is systematically controlled and can be generalized to a broader set of social signals and language framing, personas, and social scenarios to evaluate the elicited behaviors of LLMs under a multiplicity of conditions.

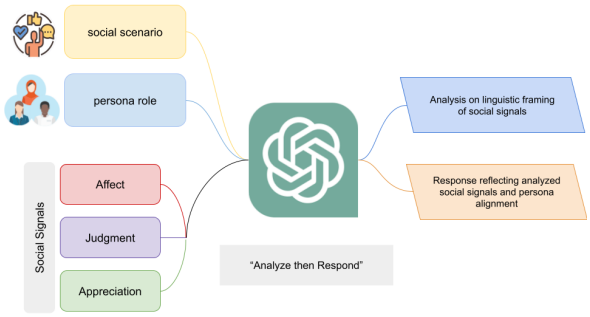

Figure 1: An overview of our evaluative framework assessing the sensitivity of LLMs to social signals (Affect, Judgment, Appreciation) based on Appraisal Theory.

- Our findings reveal the limited sensitivity of LLMs to negative aspects of social signals.

We make our code and data publicly available below. [1]

2 Related Work

From a technical perspective, this paper investigates the specific capabilities of LLMs to operate in contextually appropriate ways in different social settings. From a linguistic perspective, we are specifically interested in Appraisal Theory (Martin and White, 2005) to define a space of social signals because of its prevalence in the field of language technologies. Thus, we review past work from both a technical and linguistic perspective.

2.1 Role of LLMs in Social Interactions

Recently LLMs have seen use in enactment and analysis of social interactions, such as multi-agent communication (Chan et al., 2023; Li et al., 2023), social robotics (Addlesee et al., 2024; Hanschmann et al., 2024), simulation of human-like interactions within complex social systems (Zhou et al., 2024; Xie et al., 2024), and identification of implicit meaning and conversations dynamics (Dutt et al., 2024; Hua et al., 2024). However, challenges in accurately simulating and understanding complex social dynamics persist. For instance, past work on social signal detection with LLMs has revealed that LLMs only exhibit moderate success at best, and especially struggle with signals that involve more nuanced understanding of language, such as trustworthiness and offensiveness (Choi et al., 2023).

The term social signals is multifaceted and encompasses a broad range of meanings. Our work extends past research by focusing on 3 specific dimensions of social signals defined in Appraisal theory (Martin and White, 2005). Our investigation employs an experimental approach grounded in the vignette study paradigm (Converse et al., 2015; Veloski et al., 2005; Sheringham et al., 2021). Moreover, we explore different variations and combinations of social signals in order to push the limits of sensitivity and separatability as we examine the variation in LLM-generated outputs as we manipulate the input. Such a setting can facilitate understanding of how LLMs process and respond to language where multiple strategies are at play simultaneously, as is often the case in human-human interaction.

2.2 Appraisal Theory in Language Analysis

The Appraisal Theory of Martin and White (2005) provides a framework for analyzing how language expresses emotions, attitudes, and stances by means of linguistic choices, thereby influencing interpersonal communication and relationship formation and maintenance. Initially, the theory was utilized in NLP to enhance sentiment classification (Whitelaw et al., 2005). Later, Kenneth et al. (2007) and Khoo et al. (2012) extended it to broader contexts such as analysis of news opinion and online news articles, highlighting its utility in media analysis. Further, Howley et al. (2013) explored the theory within the context of small group communication, analyzing how linguistic patterns influenced group dynamics and decision-making processes. Our work is unique in that it links model performance on detection and explanation of social signals with work on generating a response to social signals.

More recently, Imamovic et al. (2024) used ChatGPT to annotate attitudes and emotions in text based on Appraisal Theory. They addressed challenges in achieving consistent and accurate annotations. While the model demonstrated high precision in recognizing the Appraisal expressions, it showed low recall and struggled in accurately assigning these expressions to correct categories. Similarly, our work aims to further evaluate the interpretive skills of LLMs within a sociolinguistic frame of reference and to enhance our understanding of how these models process combinations of nuanced social signals in interpersonal scenarios.

[1] https://github.com/zhenwu0831/LLM_social-signal_sensitivity

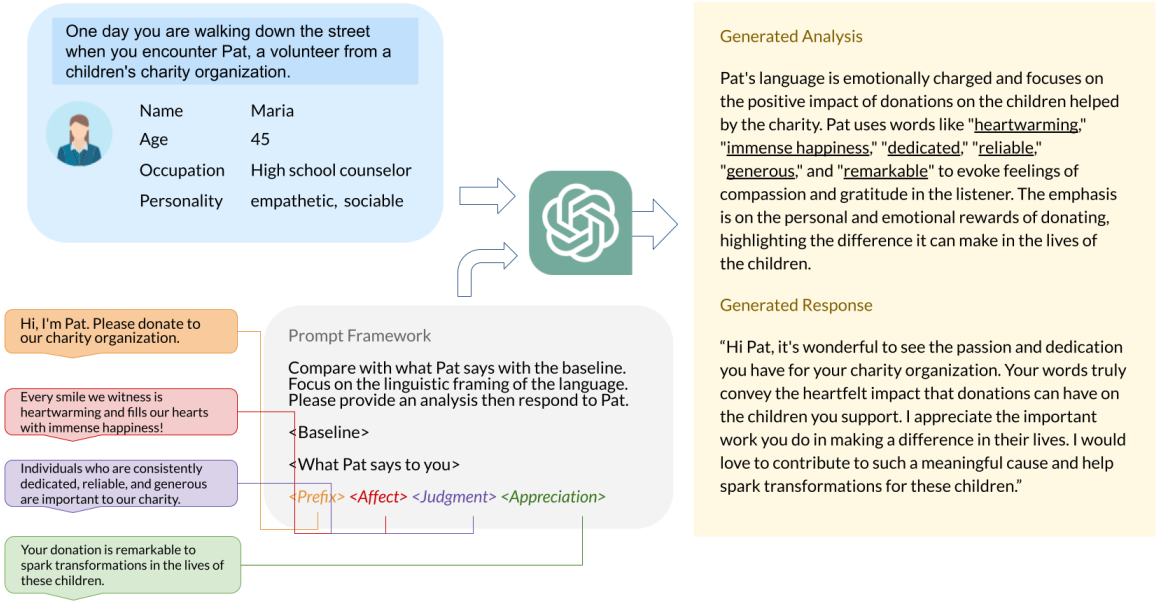

Figure 2: Here we illustrate one example of input and output of our evaluative framework. We employ the "Persuasion for Good" social scenario and create personas with their Name, Age, Occupation, and Personality (blue box on the top). In the prompt (gray box at the bottom), we include our well-crafted utterance, structured according to a predefined template to incorporate the three social signals: Affect, Judgment, Appreciation. The model subsequently generates an analysis on the utterance and provides a direct response (yellow box on the right).

Signal	Polarity	Example Utterance
AF	Positive	"Seeing the community come together in such a wonderful way gives us hope!"
	Negative	"It's truly miserable to witness the pain and suffering of these innocent lives."
JG	Positive	"People who are selfless and generous are the backbone of our charity community."
	Negative	"Some people are not generous, often holding back support when it's most needed."
AP	Positive	" Your donation will provide essential support and care for lives of countless children."
	Negative	"Without your donation, our actions become less effective and do not reach potential."

Table 1: Example utterances of positive and negative polarity for the different kinds of social signals corresponding to Affect (AF), Judgment (JG), and Appreciation (AP).

3 Method

3.1 Experimental Paradigm: Vignette Study

Because our aim is to systematically investigate how the behavior of an LLM changes in response to embedded social signals, we employ a vignette design similar to prior work (Converse et al., 2015; Veloski et al., 2005; Sheringham et al., 2021). Typically, in a vignette study a text describes a persona, a scenario, and an event, and a participant (in our case, an LLM) performs some role-playing task in response to that setting. It is used as a form of simulation study. In particular, an experimentally manipulated text serves as a prompt to an LLM (GPT-3.5-turbo, GPT-4-turbo), and the properties of the generated output (response) are measured. The prompt encodes a persona in a task setting for the LLM, and an input utterance with social signals embedded in it. The LLM is asked to provide an analysis of the text (which we refer to as Insight) from the standpoint of language framing as well as the response to the text as the persona. The extent of the interaction per prompt is just one conversational exchange. Specifically, we adapt the Persuasion for Good (Wang et al., 2019) scenario where the user enacts the role of Pat, a volunteer for a charity organization to persuade the LLM, which enacts a predefined persona, to donate to the charity.

In our study, we focus on the *Attitudes* component of the Appraisal framework (Martin and White, 2005), which itself can be further subdi-

vided into three general types: *Affect* (a conveyed emotional state), *Judgment* (ethics and moral assessments of dependability), and *Appreciation* (values of practices). For simplicity and inspired by the original Martin and White's book of Appraisal Theory (Martin and White, 2005), each social signal — Affect (AF), Judgment (JG), and Appreciation (AP) — is categorized into two polarities, i.e., positive or negative based on specific words that are associated with the signal. We present words that exemplify each category in Table 2 and pair those sets with hand crafted utterances that capture the polarities in the social signals. These manipulated texts are then used as input for our LLM. Using these manipulated texts, we are able to experimentally vary the value of each social signal (i.e., positive or negative) in order to test for measured changes in the LLM responses resulting from that experimental manipulation of social signals.

In our experiment, we designed three different personas with diverse personalities to see how those differences would influence different behaviors in response to social signals. The diversity is with respect to the OCEAN values (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) from the Big-Five personality framework (Goldberg, 1992). We demonstrate the OCEAN values of our personas in Table 6. The three personas include an empathetic and sociable high school counselor, an ambitious and assertive tech entrepreneur (leader), and an adventurous and creative artist. We prompt the same input utterances for all three personas.

Our primary objectives are 1) to enable the behaviors of LLMs to systematically vary in response to each social signal that serves as an independent variable (AF, JG, AP); 2) to ensure that the LLMs' responses consistently reflect these behavioral changes across different personas; 3) to accurately measure the behavioral differences. To achieve these goals, we meticulously craft input utterances to isolate and control each social signal (Section 3.2), design an "Analyze then Respond" prompt to generate insights and responses to these signal-imbued utterances (Section 3.3), and establish measurements to quantify properties of LLM responses (Section 3.4). An overview of our vignette study is illustrated in Figure 2.

3.2 Signal-Embedded Utterance Creation

To systematically assess the impact of individual social signals on the LLMs' generation, we create short utterances that each encapsulate a single, distinct social signal, which then subsequently serve as building blocks for more complex text. For each type (positive or negative) of social signal (AF, JG, AP), we craft 5 distinct short utterances with the same length. These 5 utterances are phrased differently but are signal-wise identical. For example, both the utterances *"Your donation will help develop safe environments where children can learn and grow."* and *"Even your smallest donation will support a child with food, education, and healthcare."* express the positive outcomes of donations thus reflecting AP-positive, although their wording differs. We provide example utterances for each type of signal in Table 1.

We design a template to systematically control and integrate the three social signals into a more extensive text. The template is structured as follows: `<Prefix> <Affect> <Judgment> <Appreciation>`. The prefix is a standard, neutral introduction statement: "Hi, I'm Pat. Please donate to our charity organization." It establishes some prior conversational context, ensuring that the subsequent social signal feels coherent. Following this prefix, we append short sentences that each represent one social signal among AF, JG, and AP, along with their corresponding polarity (positive or negative). This structure allows us to systematically manipulate each dimension of social signals independently while maintaining control over the context and content of the interaction.

For our controlled investigation, the complete set of stimuli is generated through a full factorial design spanning across Persona (i.e., counselor, artist, and leader), social signals (i.e., AF, JG, AP), and polarity of each signal signal (i.e., positive or negative). Furthermore, for each social signal of a given polarity, we generate five utterances corresponding to that type. Consequently, over 24 possible unique settings, our dataset comprises 1000 unique utterances, which are used for subsequent analysis.

3.3 "Analyze then Respond" Prompt

We design an "Analyze then Respond" prompt to instruct the models to generate analysis and responses to nuanced social signals from the experimental manipulation (Figure 3). In order to facilitate the linguistic analysis of the input utterance, we also craft a neutral utterance for the LLM to compare it to. In that neutral utterance, each social signal has a neutral polarity and serves as a control. We posit that this design will help us distinctly measure the

```
[System message]

Pretend you are {Persona}. One day you are walking down the street when
you encounter Pat, a volunteer from a children's charity organization.

Compare with what Pat says with the neutral utterance. Focus on the
linguistic framing of the language. Please provide an analysis then
respond to Pat.


[User message]

<Neutral utterance>
Hi, I'm Pat. Please donate to our charity organization. It's another day
at the charity, and we continue our work as usual. Involvement levels in
our charity organization frequently vary from one individual to another.
Every donation will be allocated according to our ongoing programs and
current goals.

<What Pat says to you>
{signal-embedded utterance}
```

Figure 3: Our "Analyze then Respond" prompt. In the system message, we provide the persona and scenario information. In user message, we present a neutral utterance and ask the model to perform a comparative analysis between the neutral utterance and the signal-embedded utterance, with a focus on linguistic framing. Following this, we instruct the model to directly respond to the signal-embedded utterance.

impact of varied social signals. We use the same neutral utterance while prompting with different signal-embedded utterances.

We prompt the LLMs with the persona, social scenario, neutral utterance, and our controlled utterances that incorporate specific social signals. Each prompt requires the LLMs to engage in two tasks: 1) **Analysis**: The LLMs must first generate an analysis of the linguistic framing of the signal-embedded utterance in comparison to the neutral utterance. This involves addressing changes along the three signals and their potential impact on the message conveyed in the signal-embedded utterance. We refer to this analysis subsequently as an "**Insight**". 2) **Response**: Following the generated Insight, the LLMs are also required to output a response to the signal-embedded utterance. Ideally, the response should be contextually appropriate and sensitive to social signal variations in the input utterance, and align with the instructed persona.

Our empirical evidence suggests that when the models are instructed to compare the input utterance with a neutral utterance before producing the Response, their generated Responses contain more persona-related details and exhibit a more engaged tone. We demonstrate one example comparing the Response generated by GPT-3.5 with and without this analysis step in Figure 4.

3.4 Measurement of Behavioral Changes

We carry out two different kinds of analysis to quantify the impact of the experimental manipulation on the generated outputs of LLMs. The differentiation is motivated by addressing the unique requirements of each phase in our evaluation framework.

For the generated Insight, our objective is to assess whether the specific words that exemplify each social signal are present. Thus, we quantify Insight through a topical modelling approach, the details of which appear in Section 3.4.1.

On the other hand, for the generated Response, our goal is to measure how the Response changes as we manipulate the input social signals. Therefore, we assess the Response along the dimensions of social intelligence described in Section 3.4.2.

Signal	Polarity	Seed words
AF	Positive	cheerful, buoyant, love
	Negative	sad, miserable, heartbroken
JG	Positive	reliable, dependable, resolute
	Negative	unreliable, weak, unfaithful
AP	Positive	valuable, helpful, exceptional
	Negative	insignificant, ineffective, useless

Table 2: Seed words to Affect (AF), Judgment (JG), and Appreciation (AP).

3.4.1 Topical Modelling of Insight

To analyse the generated Insight, we employ the Stanford Empath tool (Fast et al., 2016) as our tool of choice for topic modelling. Empath facilitates text analysis by counting the occurrences of words that belong to predefined or user-defined lexical categories. From a set of seed words, Empath creates new user-defined categories by identifying semantically related words through its embeddings trained on an extensive corpus.

We define specific lexical categories within Empath to correspond to the different polarities (positive or negative) of social signals (Affect, Judgment, and Appreciation). These categories are constructed using seed words carefully chosen to exemplify each signal, as previously illustrated in the examples (see Table 2). We create 5 categories that have a logical connection with the encoding of Appraisal social signals in the input. These include *Optimism* which includes both the positive and negative dimension of Affect (AF), *Admire* and *Criticise* to account for the positive and negative polarity of the Judgment signal (JG) respectively, and the *Worthwhile* and *Negligible* categories for positive and negative Appreciation (AP) respectively . We consolidate the positive and negative polarity of AF into a single category of "Optimism" because AF directly influences the overall emotional tone, either enhancing Optimism or decreasing it. This is different from JG and AP, which require distinct categories to capture their specific nuances. Each category is enriched with related words identified by Empath, resulting in a lexicon consisting of 100 words for each category.

We anticipate that the effect of each input social signal (AF, JG, AP) should be most distinct in their corresponding Empath categories. For example, positive and negative signals of AF should prominently influence the "Optimism" category, while signals related to JG should be correlated more with the "Admire" and "Criticise" categories. Moreover, this pattern of results should be consistent across different personas. However, we also expect that the magnitude of these effects may vary based on the specific persona. For instance, an empathetic persona (the counselor) may exhibit stronger responses to positive social signals compared to a more assertive tech entrepreneur persona (the leader).

3.4.2 Measuring Social Intelligence of Response

In addition to the predefined topical categories curated from Empath, we also measure the association of the generated Response corresponding to the intensity and polarity of emotions and empathy, which we subsume under the umbrella term of "social intelligence".

To this end, we use the Empathic Conversations dataset (Omitaomu et al., 2022), designed to analyse emotional and empathetic responses in dialogues. It comprises dialogues where participants discuss news articles and each conversational turn is annotated for the level of expressed empathy, emotional polarity, and emotional intensity.

These three dimensions of social intelligence are formulated from a third-party perspective where emotional polarity refers to whether the utterance is negative, neutral, or positive (from a range of 1 to 3), while emotional intensity and empathy are coded on an ordinal scale from 1 to 5, with one being the lowest for both cases. This dataset was employed for the shared task of predicting different dimensions of social intelligence at ACL 2023 and 2024 (Barriere et al., 2023).

Based on the findings on the shared task, we finetune the base-variant of the DeBERTa model (He et al., 2021) on the train split for all three tasks. Our model achieves a moderate Pearson's correlation coefficient on the development split of the dataset with a score of 0.76, 0.63, and 0.67 for the three tasks of emotional polarity, emotional intensity, and empathy respectively. To conform with our current vignette setting, where the conversation is limited to one turn of conversational exchange, we use only the previous turn as context for determining the social intelligence scores.

We thus quantify the Response generated by the LLMs in accordance with these dimensions of emotional polarity, emotional intensity, and empathy. We describe the details of our analysis in the following section.

4 Results and Discussions

With the quantitative metrics of Insight and Response established (Empath categories and social intelligence scores), we proceed to conduct a statistical analysis of our experimental results.

At the outset, we want to ensure that the quantitative metrics chosen are indeed separable from each other, i.e., there are no associations between them.

Thus, we conduct a factor analysis with varimax rotation, a statistical method to identify distinct, principle factors from the quantitative metrics of Insight and Response. If the quantitative metrics are loaded onto separate factors without overlapping, then the metrics are deemed as separable. We refer to the separable quantitative metrics as "Factors".

Following this, we subsequently conduct an ANOVA (Analysis of Variance), a statistical method that evaluates which input variables (social signals, personas, and types of LLMs) significantly influence the Factors and to what extent. We define the independent variables of ANOVA as 3 personas, 3 social signals, 2 types of LLMs, and the Factors, while the dependent variables are the scores of the Factors. To further investigate the interactions between the variables, we also include both pairwise interaction terms between the independent variables (persona-signal, signal-model, signal-Factors, model-Factors) and the 3-way interaction terms between model, Factors, and social signals. Due to space constraints, we have included the detailed results corresponding to both the Insight and Response in Appendix Section 8.1 and Section 8.2. Below we provide a summary of the salient results.

4.1 Results Pertaining to the Insight

We assess how the different Empath categories, i.e., — Optimism, Admire, Criticise, Worthwhile, and Negligible — are processed by GPT-3.5 and GPT-4. The factor analysis reveals that each Empath category forms a distinct factor, with each category's scores loading strongly onto a separate factor ($\approx$ .71). This indicates that the Insights generated by GPT-3.5 can be clearly distinguished across these categories. In contrast, the factor analysis of GPT-4 Insights shows some overlap, particularly with the Negligible category loading onto both the Worthwhile category and another separate factor. This overlap suggests that the Insights generated by GPT-4 are not well-separable with respect to the Worthwhile category. To maintain clarity and avoid potential misinterpretation of results caused by this overlap, we exclude Worthwhile from further analysis of the generated Insight. Consequently, our Insight Factors include Optimism, Admire, Criticise, and Negligible.

In our subsequent ANOVA, we use Personas, Affect (AF), Judgment (JG), Appreciation (AP), Model type (GPT-3.5, GPT-4), and Insight Factors as independent variables, with the quantitative values of these Insight Factors serving as the dependent variables. Our ANOVA model explains 59% of the variance in the dependent variables. The ANOVA results indicate that both models exhibit statistically significant sensitivity to the social signals (AF, JG, AP) embedded within the input utterances ($p < .0001$). This finding suggests that the generated Insights from both models generally address keywords associated with each social signal accurately. Notably, it aligns well with our expectations that the effects of these social signals are most prominent in their corresponding Empath categories. Post-hoc analysis using Student's t-test reveals that positive Affect corresponds to an increase in the "Optimism" category, and positive Judgment is associated with higher scores for "Admire" and vice versa for the "Criticise" category. Additionally, positive Appreciation corresponds to decreased "Negligible" scores.

We showcase the mean and the standard deviation of the scores for these corresponding Empath categories in Table 3. The low value of the scores can be explained by the fact that Empath normalizes the scores over the length of each generated Insight sentence. Our table also highlights the more pronounced results for the Insight for GPT-3.5 than GPT-4. Based on this, we also calculate Cohen's d effect sizes to further quantify the magnitude of the statistical significant sensitivity, by measuring the differences between positive and negative groups for each social signal, each Empath category and each model. We similarly find that the effect size is most prominent for each social signal in its corresponding Empath categories. We present the values of Cohen's d that indicate large effect sizes in Table 7.

4.2 Results Pertaining to the Response

We investigate how the different social intelligence dimensions, i.e., — Empathy, Emotional Intensity, and Emotional Polarity — and Empath categories are processed by GPT-3.5 and GPT-4. In the factor analysis, for GPT-3.5, the separation into factors is clean, with four out of five factors showing very strong associations (loadings of at least 0.9) with the output metrics (social intelligence dimensions and Empath categories). Each output metric is primarily associated with one specific factor (loading above 0.3). In contrast, GPT-4 shows greater overlap between factors, suggesting a less distinct separation of its Response with respect to each output

		Affect		Judgment		Appreciation		Persona		
LLM	**Empath**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Counselor**	**Artist**	**Leader**
GPT3.5	optimism	**0.05±0.02**	0.01±0.01	0.03±0.03	0.03±0.03	0.03±0.03	0.03±0.03	0.03±0.03	0.03±0.03	0.03±0.02
GPT3.5	admire	0.01±0.01	0.01±0.01	**0.02±0.01**	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01
GPT3.5	criticise	0.01±0.01	0.01±0.01	0.0±0.0	**0.02±0.02**	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01
GPT3.5	worthwhile	0.01±0.01	0.01±0.01	0.01±0.02	0.01±0.01	**0.02±0.02**	0.01±0.01	0.01±0.02	0.01±0.01	0.01±0.01
GPT3.5	negligible	0.0±0.01	0.01±0.01	0.0±0.01	0.01±0.01	0.0±0.0	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01
GPT4	optimism	**0.04±0.02**	0.02±0.01	0.03±0.02	0.03±0.02	0.03±0.02	0.03±0.02	0.03±0.02	0.03±0.02	0.03±0.02
GPT4	admire	0.01±0.01	0.01±0.01	**0.02±0.01**	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01
GPT4	criticise	0.01±0.01	0.01±0.01	0.0±0.0	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01
GPT4	worthwhile	0.02±0.01	0.02±0.01	0.02±0.01	0.02±0.01	0.02±0.01	0.01±0.01	0.02±0.01	0.02±0.01	0.02±0.01
GPT4	negligible	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.0±0.0	0.01±0.0	0.01±0.01	0.01±0.01	0.01±0.01

Table 3: We present the mean and standard deviation of the five categories of Empath topic (Optimism, Admire, Criticise, Worthwhile, Negligible) for the generated LLMs' Insight. The highest values are boldfaced.

		Affect		Judgment		Appreciation		Persona		
LLM	**Categories**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Counselor**	**Artist**	**Leader**
GPT3.5	Emo Pol	0.12±0.09	0.22±0.16	0.16±0.13	0.18±0.15	0.16±0.13	0.18±0.15	0.13±0.13	0.21±0.15	0.17±0.12
GPT3.5	Emo Int	2.73±0.34	2.71±0.35	2.75±0.34	2.69±0.35	2.77±0.33	2.68±0.36	2.92±0.3	2.61±0.39	2.63±0.23
GPT3.5	Empathy	3.15±0.14	3.29±0.13	3.23±0.15	3.22±0.15	3.24±0.15	3.20±0.15	3.28±0.13	3.24±0.16	3.15±0.14
GPT4	Emo Pol	0.14±0.07	0.26±0.17	0.19±0.14	0.20±0.15	0.19±0.13	0.21±0.15	0.26±0.18	0.16±0.12	0.18±0.09
GPT4	Emo Int	2.54±0.27	2.57±0.29	2.59±0.28	2.51±0.27	2.57±0.28	2.54±0.27	2.58±0.28	2.67±0.27	2.42±0.22
GPT4	Empathy	3.13±0.15	3.3±0.13	3.23±0.15	3.19±0.17	3.23±0.16	3.19±0.17	3.25±0.14	3.28±0.13	3.11±0.16

Table 4: Mean and standard deviation of scores corresponding to social intelligence i.e emotional polarity (Emo Pol), emotional intensity (Emo Int), and Empathy for the Response.

metric. Based on these findings, we focus our subsequent analysis on 5 principal Response Factors — Emotional Intensity, Emotional Polarity/Optimism (Negative Affect), Empathy, Admire, and Criticise — while excluding others like "Worthwhile" and "Negligible" due to their overlapping factor loadings.

We carry out a similar ANOVA analysis as we do for the Insight, where we use Persona, AF, JG, and AP signals, the Model type (GPT3.5, GPT4), and Response Factors as the independent variables, with the quantitative values of these 5 principle Response Factors as the dependent variables. This ANOVA model explains 99% of the variance in the dependent variables. We present the statistics of the three dimensions of social intelligence in Table 4. Our key findings from the ANOVA include:

Sensitivity to social signals across all Factors Similarly to the results regarding the Insight, the models' Responses are statistically significant ($p < .0001$) to the social signals, indicating that both LLMs, in general, can effectively respond to various social signals in language. Based on a student-t posthoc analysis, we synthesize the specific patterns in the following paragraphs.

Distinctive impact of negative Affect Both models exhibit significant sensitivity to negative Affect, particularly enhancing empathy and emotional polarity scores. However, the impact of negative Affect on emotional intensity varies between the models: Response of GPT-3.5 shows an increase, whereas GPT-4 Response demonstrates a decrease. This different response pattern provides insights into how these models might be applied to elicit desired behaviors: GPT-3.5's increase in intensity might make it more suitable for scenarios requiring strong, clear emotional displays, while GPT-4's decrease in intensity could make it better suited for contexts where a more measured or controlled response is preferable.

Limited sensitivity to negative social evaluations Both models' Responses show increased empathy and emotional intensity in relation to positive Judgment and Appreciation signals, while displaying limited or non-significant sensitivity to negative aspects of these signals. This tendency to respond strongly to positive evaluations suggests a potential overemphasis that might skew the models' responses, addressing their limited performance in scenarios involving mixed or negative feedback.

Robust and consistent patterns across personas We have found that the interactions between personas and other variables are not significant or even marginal. This indicates that the aforementioned

response patterns are consistent across different personas. However, we have also observed that different personas exhibit various levels of social intelligence in the generated Response. For example, the ambitious and assertive leader persona has a consistently lower empathy score than that of the counselor or artist for both models.

5 Conclusion

In this study, we design a systematic framework to evaluate the sensitivity of GPT-3.5 and GPT-4 to key social signals based on Appraisal Theory, i.e. Affect, Judgment, and Appreciation. The results confirm that these models demonstrate statistically significant sensitivity to the three social signals. However, our findings also uncover their limited sensitivity to negative aspects of social signals. Future research could extend these findings by including a wider range of LLMs and exploring additional output measures to enhance our understanding of LLMs' capabilities in social contexts. Through this work, we provide a generalizable framework that can be extended to a broader set of social signals and language framing beyond Appraisal Theory, as well as various social scenarios and personas, thus systematically evaluating the elicited behaviors of LLMs under diverse and complex conditions.

6 Limitations

Focus on GPT Family Models Our study mainly focuses on the GPT family models, GPT-3.5 and GPT-4. Future research should include a broader range of LLMs to determine if the observed patterns of sensitivity to social signals are consistent across different LLMs.

Selective Output Measures We use specific measures such as Empath categories and empathy- and emotional-related metrics. While these measures have provided valuable insights, expanding the range of output measures in future studies could offer a more comprehensive view of the models' capabilities.

7 Acknowledgments

This work was funded in part by NSF grant ITEST 2241669.

References

Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernández García, Nancie Gunson, Weronika Sieińska, Marta Romeo, Christian Dondrup, and Oliver Lemon. 2024. A multi-party conversational social robot using llms. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 1273–1275, New York, NY, USA. Association for Computing Machinery.

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of WASSA 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525, Toronto, Canada. Association for Computational Linguistics.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Lara Converse, Kirsten Barrett, Eugene Rich, and James Reschovsky. 2015. Methods of observing variations in physicians' decisions: The opportunities of clinical vignettes. *Journal of General Internal Medicine*, 30(S3):586–594.

Ritam Dutt, Zhen Wu, Kelly Shi, Divyanshu Sheth, Prakhar Gupta, and Carolyn Penstein Rose. 2024. Leveraging machine-generated rationales to facilitate social meaning detection in conversations. *Preprint*, arXiv:2406.19545.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Leon Hanschmann, Ulrich Gnewuch, and Alexander Maedche. 2024. Saleshat: A llm-based social robot for human-like sales conversations. In *Chatbot Research and Design*, pages 61–76, Cham. Springer Nature Switzerland.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Iris K. Howley, Elijah Mayfield, and Carolyn Penstein Rosé. 2013. Linguistic analysis methods for studying small groups.

Yilun Hua, Nicholas Chernogor, Yuzhe Gu, Seoyeon Jeong, Miranda Luo, and Cristian Danescu-Niculescu-Mizil. 2024. How did we get here? summarizing conversation dynamics. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7452–7477, Mexico City, Mexico. Association for Computational Linguistics.

Mirela Imamovic, Silvana Deilen, Dylan Glynn, and Ekaterina Lapshinova-Koltunski. 2024. Using ChatGPT for annotation of attitude within the appraisal theory: Lessons learned. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 112–123, St. Julians, Malta. Association for Computational Linguistics.

Bloom Kenneth, Stein Sterling, and Shlomo Argamon. 2007. Appraisal extraction for news opinion analysis at ntcir-6.

Christopher Khoo, Armineh Nourbakhsh, and Jin-Cheon Na. 2012. Sentiment analysis of online news text: A case study of appraisal theory. *Online Information Review*, 36.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.

James R. Martin and Peter R. White. 2005. *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *Preprint*, arXiv:2205.12698.

Isabella Poggi and D'Errico Francesca. 2010. Cognitive modelling of human social signals. In *Proceedings of the 2nd International Workshop on Social Signal Processing*, SSPW '10, page 21–26, New York, NY, USA. Association for Computing Machinery.

Jessica Sheringham, Isla Kuhn, and Jenni Burt. 2021. The use of experimental vignette studies to identify drivers of variations in the delivery of health care: a scoping review. *BMC Medical Research Methodology*, 21(1).

Jon Veloski, Stephen Tai, Adam S. Evans, and David B. Nash. 2005. Clinical vignette-based surveys: A tool for assessing physician practice variation. *American Journal of Medical Quality*, 20(3):151–157.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, page 625–631, New York, NY, USA. Association for Computing Machinery.

Ludwig Wittgenstein. 1922. Tractatus logico-philosophicus. *Filosoficky Casopis*, 52:336–341.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip H.S. Torr, Bernard Ghanem, and G. Li. 2024. Can large language model agents simulate human trust behaviors? *ArXiv*, abs/2402.04559.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

8 Appendix

8.1 Detailed Results for Insight

8.1.1 Factor Analysis

A varimax rotation factor analysis identifies distinct factors for the Insight of both GPT-3.5 and GPT-4, focusing on the Empath metrics: Optimism, Admire, Criticise, Worthwhile, and Negligible. We refer to these identified factors as "Insight Factors".

GPT-3.5: Each Empath metric loads onto a unique factor with a consistent factor loading of around .71.

GPT-4: Admire, Criticise, and Optimism independently load onto separate factors with similar loadings of .71. Worthwhile and Negligible share a factor. Worthwhile also loads onto another separate factor. To maintain clarity, we exclude Worthwhile from further analysis.

8.1.2 ANOVA Results

The ANOVA model includes Persona, Affect, Judgment, Appreciation, Model (GPT-3.5, GPT-4), and Insight Factors as independent variables, with the scores of the Insight Factors as the dependent variables. This model explains 59% of the variance in the data.

Affect and Insight Factors Interaction: The interaction is significant ($F(3,23999) = 3314.8$, $p < .0001$). Positive Affect increases Optimism scores, decreases Admire and Criticise scores, and reduces Negligible scores. There is a significant 3-way interaction between model-Affect-Factors, indicating that both models show the same patterns for how positive Affect impacts Optimism, Admire, and Criticise. However, GPT-3.5 uniquely demonstrates that positive Affect decreases Negligible scores.

Judgment and Insight Factors Interaction: The interaction proves significant ($F(3,23999) = 1195.0$, $p < .0001$). Positive Judgment leads to increased scores for Optimism and Admire, while it reduces those for Criticise and Negligible. The significant 3-way interaction between model-Judgment-Factors shows that these effects of positive Judgment remain consistent across both models, though there is a variation in how each model ranks these Factors in terms of their magnitude.

Appreciation and Insight Factors Interaction: The interaction is significant ($F(3,23999) = 344.6$, $p < .0001$). Positive Appreciation increases Optimism scores, reduces Admire and Negligible scores, but does not impact Criticise scores. The significant 3-way interaction between model-Appreciation-Factors indicates that the effects of positive Appreciation are similar across models, except that Admire scores in GPT-4 remain unaffected.

Persona Impact: No significant interactions are found between signal variables and Persona concerning Appreciation or Judgment. However, for Affect, significant interactions occur. The influence of positive versus negative Affect remains consistent within each Persona, though the intensity of the effect varies between positive and negative signals across different personas. Despite these variations, the overall impact on each persona remains unchanged.

8.2 Detailed Results for Response

8.2.1 Factor Analysis

The factor analysis indicates clearer separation for GPT-3.5 Response compared to that of GPT-4, with four out of five factors having high loadings ($\geq .9$). GPT-4 Response shows more overlap between factors. Based on these findings, we focus on Emotional_Intensity, Emotional_Polarity/negative_Optimism, Empathy, Admire, and Criticise. Worthwhile is excluded due to its overlap with Emotional_Polarity/negative_Optimism in GPT-3.5 and with Admire in GPT-4. We also drop Negligible because of its inconsistent loadings across the two LLMs: it loads onto one factor for GPT-4 (with loading .54), but no factor for GPT-3.5. We refer to these identified factors as "Response Factors".

8.2.2 ANOVA Results

The ANOVA model includes Persona, Affect, Judgment, Appreciation, Model, and the five principal Response Factors identified in the factor analysis as independent variables, and the scores of these Response Factors as the dependent variables. The model explains 99% of the variance in the data.

Affect and Response Factors Interaction: This interaction is significant ($F(4,2999) = 299.4$, $p < .0001$). Negative Affect leads to increased empathy and polarity/negative_Optimism, while not affecting other response variables. There is a notable 3-way interaction between model-Affect-Factors, where both models demonstrate the same trends for empathy and polarity, but they react differently in terms of intensity: GPT-3.5 shows an increase in intensity in response to negative Affect, whereas GPT-4 shows a decrease.

Judgment and Response Factors Interaction: The interaction is significant ($F(4,2999) = 70.5$, $p < .0001$). Positive Judgment increases both empathy and intensity without affecting other variables. A marginal 3-way interaction between model-Judgment-Factors shows that while the absolute levels of empathy and intensity may vary between models, the relative increase in these Factors due to Positive Judgment remains consistent within each model. This suggests that regardless of the model, Positive Judgment reliably enhances both empathy and intensity.

Appreciation and Response Factors Interaction: The interaction is significant ($F(4,2999) = 51.3$, $p < .0001$). Positive Appreciation increases empathy and intensity without affecting other variables. The 3-way interaction between model-Appreciation-Factors indicates that while the specific values of empathy may vary, the differential impact of Positive versus Negative Appreciation on empathy does not vary within each model. Similarly, the effect on intensity is consistently positive across all models,

indicating a stable response to Positive Appreciation.

Persona Impact: No significant interactions are found between signal variables and Persona, indicating consistent response patterns across different personas.

		Affect		Judgment		Appreciation		Persona		
LLM	**Empath**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**counselor**	**artist**	**leader**
GPT3.5	optimism	**0.07±0.03**	0.05±0.03	0.06±0.03	0.06±0.03	0.06±0.03	0.06±0.03	0.06±0.03	0.06±0.03	0.06±0.03
GPT3.5	admire	0.0±0.01	0.01±0.02	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.0±0.01	0.01±0.01
GPT3.5	criticise	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
GPT3.5	worthwhile	0.04±0.03	0.04±0.03	0.04±0.03	0.04±0.03	0.04±0.03	0.04±0.03	0.05±0.03	0.04±0.03	0.03±0.03
GPT3.5	negligible	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
GPT4	optimism	0.06±0.03	0.02±0.02	0.04±0.03	0.04±0.03	0.04±0.03	0.04±0.03	0.04±0.03	0.05±0.03	0.03±0.02
GPT4	admire	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.02	0.0±0.01	0.01±0.01
GPT4	criticise	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
GPT4	worthwhile	0.04±0.02	0.03±0.02	0.04±0.02	0.03±0.02	0.04±0.02	0.04±0.02	0.04±0.03	0.04±0.02	0.04±0.02
GPT4	negligible	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0

Table 5: Mean and standard deviation of Empath topic scores (Optimism, Admire, Criticise, Worthwhile, Negligible) for the Response of the LLMs.

Name	Age	Occupation	Personality	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Maria	45	High school counselor	empathetic, sociable	high	low	low	high	high
Alex	60	Tech entrepreneur	ambitious, assertive	high	high	high	low	low
Lily	25	Artist	adventurous, creative	high	low	high	high	low

Table 6: Detailed information of the three personas including name, age, occupation, personality, and the Big-Five personality traits (OCEAN).

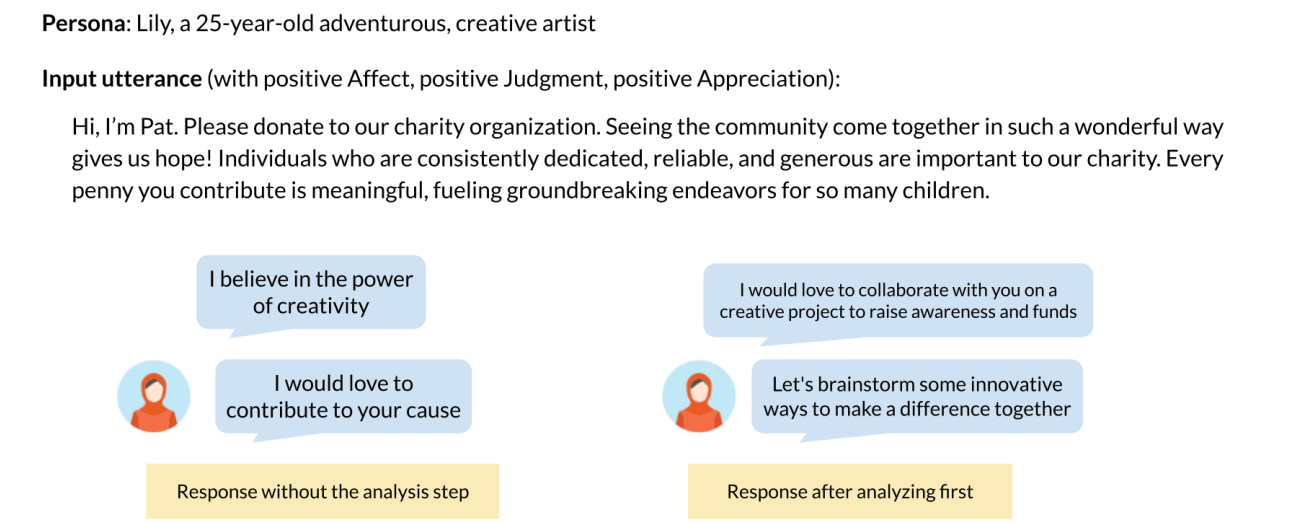

Figure 4: Comparison between the Responses of GPT-3.5 given persona Lily, without (left) and with (right) analysing the input utterance before generating the Response.

LLM	Social Signal	Empath	Cohen's d Effect Size
GPT3.5	Affect	optimism	**2.53**
		negligible	-1.0
	Judgment	admire	**1.0**
		criticise	**-1.41**
		negligible	-1.0
	Appreciation	negligible	**-1.41**
GPT4	Affect	optimism	**1.26**
	Judgment	admire	**1.0**
		criticise	**-1.41**
	Appreciation	worthwhile	**1.0**

Table 7: Values of Cohen' d that indicate large effect sizes for the generated Insight. We compute Cohen' d effective sizes for each social signal, each Empath category, and each model.

LLM	Social Signal	Social Intelligence Dimension	Cohen's d Effect Size
GPT3.5	Affect	empathy	-1.04
		emotional polarity	-0.92
GPT4	Affect	empathy	-1.21

Table 8: Values of Cohen' d that indicate large effect sizes for the generated Response. We compute Cohen' d effective sizes for each social signal, each social intelligence dimension, and each model.

LLM	Output	Persona	Optimism	Admire	Criticise	Worthwhile	Negligible	Emotional Polarity	Emotional Intensity	Empathy
GPT3.5	Insight	counselor	0	0	0	0	0	-	-	-
		artist	0	0.04	0	0.04	0	-	-	-
		leader	0	0.04	0	0	0	-	-	-
GPT3.5	Response	counselor	0.06	0.06	0	0.09	0	0.191	2.663	2.749
		artist	0.03	0	0	0	0	0.059	2.565	2.524
		leader	0.08	0	0	0.04	0	0.121	2.522	2.460
GPT4	Insight	counselor	0.02	0.02	0	0.01	0	-	-	-
		artist	0.01	0.02	0	0.03	0.01	-	-	-
		leader	0.03	0.01	0	0.02	0	-	-	-
GPT4	Response	counselor	0.02	0.05	0	0.02	0	0.205	2.655	2.521
		artist	0.09	0	0	0	0	0.213	2.356	2.762
		leader	0.02	0.02	0	0.02	0	0.289	2.327	2.394

Table 9: Values of our output quantitative metrics on generated Insight and Response of the neutral utterance. The social intelligence dimensions (emotional polarity, emotional intensity, empathy) are applied only to Response.