

SocialLM: Social Signal Processing of Patient-Provider Communication using LLMs and Contextual Aggregation

Manas Satish Bedmutha
UC San Diego, USA

MBEDMUTHA@UCSD.EDU

Feng Chen
University of Washington, USA

FENG9@UW.EDU

Andrea L Hartzler
University of Washington, USA

ANDREAH@UW.EDU

Trevor Cohen
University of Washington, USA

COHENTA@UW.EDU

Nadir Weibel
UC San Diego, USA

WEIBEL@UCSD.EDU

Abstract

Effective patient–provider communication is difficult to assess at scale. We examine whether large language models (LLMs) can track 20 social behaviors from clinical transcripts without fine-tuning. Across three model families and multiple prompting strategies, LLMs reliably detect social signals, though performance varies by patient race and visit segment. To address this variability under query-only API constraints, we introduce an agreement-weighted ensemble using group-level agreement patterns. This approach improves both accuracy and stability over the best individual model, demonstrating a practical pathway for scalable social signal tracking in clinical conversations.

Data and Code Availability We use conversations from the *Establishing Focus* (EF) dataset (AHRQ, 2006), which cannot be publicly shared due to the usage of real-world recordings of patient-provider conversations in clinic visits; interested researchers may contact the authors regarding data-use agreements. Code available [here].

Institutional Review Board (IRB) This secondary analysis is approved by the IRBs of collaborating institutions and is HIPAA-compliant (UW STUDY00005436).

1. Introduction

Communication between patients and primary-care providers (henceforth *providers*) is critical to patient wellbeing. Good communication improves trust, adherence and visit-satisfaction, thereby improving health outcomes (Street Jr et al. (2007)). Clinical conversations are different from everyday conversations due to inherent power asymmetries, time constraints and high stakes, thereby making even subtle interpersonal dynamics consequential. As a result, there is growing interest in tools that help providers reflect on and improve their communication behaviors, such as the ones developed by Wang et al. (2024) and Bedmutha et al. (2024b).

Prior work has examined these interpersonal behaviors through the *social signals*. Vinciarelli et al. (2009) defines them as *aggregates of verbal and non-verbal behaviors that individuals exhibit towards others within their current social context*. In clinical settings, the most prominent set of behaviors is the Roter Interaction Analysis System (RIAS) (Roter and Larson (2002)), which tracks patients and providers on a set of 21 signals. To obtain fine-grained understanding of the clinical conversations, they recommended breaking visits into *thin slices* of shorter durations (typically 3-5 minutes) and annotate each “slice” for all social signals. This requires *expert annotators trained in RIAS* and is time-intensive, therefore **cannot provide communication feedback at scale**.

Empirically most social signals can be characterized through non-verbal cues such as prosodics and turn-taking (Hartzler et al. (2014)). Bedmutha et al. (2024b) aimed to leverage this association to train models that can track the RIAS signals using such vocalic cues. Despite annotating numerous recordings, they found that the prevalence of high/low levels of social signals was very skewed and in fact they could only model 8 social signals out of the 21. The imbalance is systemic and occurs due to Hawthorne effect described by Landsberger (1958), and may not be ethically addressable with additional data. Chen et al. (2024a) explored the use of text or content of the conversations to fine-tune generalized small language models (SLMs). Despite the limitations, they trained models for each individual task and showed feasibility of tracking 17 out of the 21 signals evaluated across stratified groups of providers. Since the cost of annotating new data is extremely high and the data is highly imbalanced, we envision that tracking these social behaviors requires any model to have **inherent knowledge** of social and affective behaviors.

While non-verbal cues may offer actionable feedback, their meaning significantly differs across cultures, and audio and video may often be unavailable in real-world deployments. In contrast, historically transcripts have been generated and stored using manual systems. Recently, automated transcription is rapidly becoming standard through ambient clinical intelligence systems, and as a result, large volumes of clinical conversation transcripts are generated and available. As shown by Swain (2024), transcripts also offer stronger privacy protections compared to raw audio data. This has created an opportunity to infer interpersonal dynamics directly from text alone.

Xu et al. (2024) and Yang et al. (2023) showed how large language models have demonstrated strong capabilities in extracting clinically relevant information from free text, including mental health states such as depression and anxiety. Emerging evidence by Chen et al. (2025) and Montiel-Vázquez et al. (2024) suggests that language models may also capture interpersonal and affective behaviors in other related settings such as therapy. However, it remains unclear whether such models can robustly infer a broad range of social signals, if the performance varies based on known contextual indicators, and whether these predictions can be made reliable enough for deployment.

In this work, we study whether large language models (LLMs) exhibit *inherent capability* to infer so-

cial signals from clinical transcripts without supervised training. We evaluate three language models spanning a wide range of scales and architectures, from a small on-device model to a large cloud-hosted model. Across 20 social signals annotated in three-minute transcript segments, we assess zero-shot and few-shot prompting strategies and analyze performance variability across visit stages and patient demographics. Finally, inspired from real-world deployments where query-only models are encouraged and hence uncertainty and confidence scores are unavailable; we investigate the impact of ensembling different model-prompt responses for each task by introducing context-group level agreement probabilities as soft weights.

Our results show that language models can act as scalable text-based sensors of interpersonal dynamics in clinical conversations, offering a practical pathway for monitoring and studying social signals in real-world healthcare settings.

2. Related Work

Our work builds on prior research on patient-provider communication, social signal processing and the use of large language models in similar tasks.

2.1. Social Signals in Clinical Conversations

Patient-provider communication has been widely studied, mostly empirically by Cooper et al. (2012), Roter and Larson (2002), and Hartzler et al. (2014). Roter and Larson (2002) define 21 social signals through the Roter Interaction Analysis System (RIAS) that impact the success of a visit: *dominance, *attentiveness, *warmth, *engagement, *empathy, *respect, *interactivity, *irritation, *nervousness, **hurriedness, ***sadness, and **emotional distress (*patient and provider, **provider only, ***patient only). Recent work by Chen et al. (2024a), Bedmutha et al. (2023b), and Bedmutha et al. (2024a) has shown that RIAS-inspired social signals may present differently based on patients' race. Additionally, since patient visits often follow a structured flow as described by Robinson (2003), social signal distributions may vary at the start, middle, and end of the visit. These factors motivate examining whether AI performance differs by race and time-within-visit.

Early computational approaches modeled social signals via acoustic, prosodic, and turn-taking features. Hartzler et al. (2014) and Liu et al. (2016)

used audio/video, while [Bedmutha et al. \(2024b\)](#) and [Wang et al. \(2024\)](#) focused on vocalic cues—but these typically addressed few signals and required high-quality recordings. Text-based approaches have gained attention with increasing availability of clinical transcripts; [Chen et al. \(2024a\)](#) showed that fine-tuned small language models could predict up to seventeen RIAS signals. However, these methods rely on supervised, task-specific training, limiting scalability. To date, no prior work has examined whether LLMs, as foundation models without fine-tuning, can infer a broad range of social signals from text alone.

Building reliable systems remains challenging due to labor-intensive annotation, signal sparsity, and limited generalizability of institution-specific models. Yet the growing ubiquity of AI scribes, as noted by [Tierney et al. \(2024\)](#), and privacy-aware sensing systems, such as the ones developed by [Bedmutha et al. \(2023a\)](#) and [Boovaraghavan et al. \(2024\)](#) suggest transcripts will soon be widely available. [Bedmutha et al. \(2024a\)](#) and [Bascom et al. \(2024\)](#), have emphasized that providers value timely, personalized feedback, pointing to the need for unified systems that operate without extensive retraining.

2.2. Large Language Models for Patient-Provider Communication

In healthcare, LLMs have shown promise for clinical documentation and mental health assessment, as demonstrated by [Xu et al. \(2024\)](#) and [Yang et al. \(2023\)](#). For patient-provider communication specifically, [Ayers et al. \(2023\)](#) found that LLM-generated responses can be perceived as accurate and empathic in asynchronous settings, while [Vishwanath et al. \(2024\)](#) and [Tu et al. \(2025\)](#) have explored LLMs for patient education and conversational agents.

Recent studies suggest LLMs can reason about affective properties of health-related text. [Luo et al. \(2024\)](#) showed LLMs can assess empathy in patient-provider messages, and [Feng et al. \(2023\)](#) demonstrated emotional state capture in therapeutic conversations; outside healthcare, [Lei et al. \(2023\)](#) found strong performance on emotion recognition in dialogue.

However, it remains unclear whether these capabilities extend to naturalistic, synchronous clinical conversations involving role asymmetries and subtle social cues; or whether performance varies across patient populations, visit stages, and signal types.

3. Modeling Social Behavior in Clinical Conversations

We frame social signal processing (SSP) as the task of inferring interpersonal behaviors expressed in short segments of clinical conversations or slices. Because many social behaviors lack precise definitions and exhibit substantial annotator subjectivity, SSP can also be interpreted as a human-AI alignment problem, where a correct prediction reflects agreement with expert human judgments rather than an objective ground truth.

3.1. Dataset and Task

We evaluate SSP using the Establishing Focus (EF) dataset from [AHRQ \(2006\)](#), which contains real-world primary care visits from the West Coast of the United States, annotated with RIAS-inspired global affect ratings as defined by [Roter and Larson \(2002\)](#). This is the largest annotated dataset and used widely for empirical (such as [Vannoy and Robins \(2011\)](#); [Bascom et al. \(2024\)](#)) and quantitative studies (such as [Bedmutha et al. \(2024b\)](#); [Zolensky et al. \(2025\)](#)).

Following the thin-slice methodology of [Roter et al. \(2011\)](#), each visit is segmented into independent three-minute “thin slices”, yielding 512 annotated segments across 91 visits conducted by 22 providers. The audio recordings are transcribed using Whisper-v3-large ([Radford et al., 2022](#)) and diarized using pyannote ([Bredin et al., 2020](#)), which have become the standard for conversational affective tasks [Agrawal et al. \(2025\)](#). After preprocessing for empty slices, our final dataset consists of 508 high-quality slices. The coding scheme includes 21 social signals attributed to the patient, provider, or both. These are grouped into:

1. *Type-I Signals*: Dominance, attentiveness, warmth, engagement, empathy, respect, interactivity, and hurriedness. All are rated for both patient and provider except hurriedness (provider only). Scores range from 1 (low) to 6 (high), with 3.5 as the neutral midpoint.
2. *Type-II Signals*: Irritation, nervousness, sadness, and emotional distress; sadness and emotional distress are patient-only. These capture negative affective states, where 1 denotes absence and ratings increase with expression up to 6.

One signal (provider nervousness) is never observed, leaving 20 signals for analysis. Following

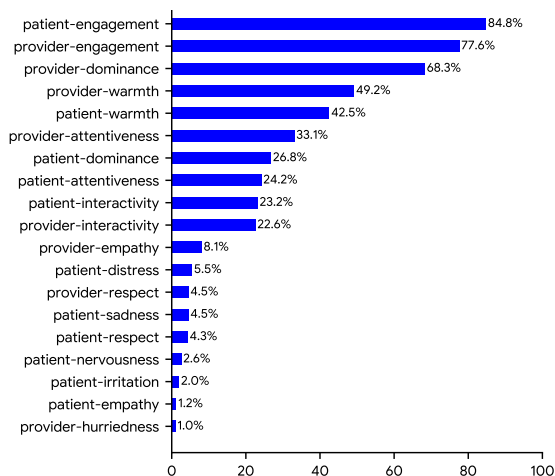


Figure 1: Percentage of *high* social signal labels distribution across 3-minute segments across the entire sample

prior work by [Chen et al. \(2024a\)](#) and [Bedmutha et al. \(2024b\)](#), we binarize all labels. For Type-I signals, we predict whether behavior is expressed at a higher-than-average level (threshold 3.5). For Type-II signals, we predict presence versus absence (threshold 1.5). This formulation reflects both limited label granularity and the extreme sparsity of negative or disruptive behaviors in clinical encounters. The final data distribution ([Figure 1](#)) yields a binary prediction task across 20 behaviors per 3-minute slice of conversation.

Present-day utility: [Bedmutha et al. \(2024a\)](#) collected another dataset of 108 visits in 2023 consisting of a diverse representation of LGBTQ+ and BIPOC participants. They trained social signal models on the EF dataset using primitive audio features and showed that these models directly generalized on their new dataset in 5/7 signals and all models beat the fairness baselines. This instills confidence that findings from our current analysis have the potential to translate to present-day communication patterns.

3.2. Evaluating Feasibility of SSP

We describe below the overview of our feasibility benchmarking to explore if at all language models can track affective information from clinical conversations. The EF dataset exhibits severe class imbalance across nearly all signals ([Figure 1](#)), particularly for Type-II behaviors, which are rare or nearly absent. Moreover, positive instances are often clustered within a small number of visits or providers, limit-

ing diversity and making supervised learning unstable. These challenges stem from known biases in observational healthcare data, including sampling bias and the Hawthorne effect. As a result, training task-specific models from scratch or via fine-tuning is infeasible for many signals. Instead, we investigate whether large language models can leverage internalized social and conversational knowledge to perform SSP in zero- and few-shot settings, without relying on extensive labeled data.

3.2.1. PROMPTING AND MODELS

Each social signal is treated as an independent binary classification task applied to each transcript slice. We evaluate standard prompting strategies including zero-shot, few-shot, and chain-of-thought prompting. Prompts specify the model’s role, the target social signal, and the required output format. Further details of the prompts are provided in [Appendix A](#).

We evaluate three representative language models spanning a wide range of scales: a small instruction-tuned encoder-decoder model (Flan-T5-Base), a mid-sized decoder-only model (Gemma2-2B), and a large-scale decoder-only model (Llama 3.1 405B). All models are run locally to ensure data privacy. Flan and Gemma are used in *generation* while Llama is used in *instruct* mode.

Clinical Feasibility: The first two models represent on-device deployment; the latter simulates cloud inference. Our health system’s protocols restricted access to logit distributions, so we operate under the constraint that large models are **query-access only**. Despite the binarized state, tracking changes in these behaviors can support various reflective visualizations for providers [Bedmutha et al. \(2024b\)](#); ?

3.2.2. EVALUATION

Each `<model, prompt>` pair is treated as a distinct configuration. Models are run deterministically using greedy decoding (temperature zero), ensuring consistent responses and removing the need for resampled evaluations. When applicable, class predictions are derived via logit comparison between binary labels. Some prompt-model combinations were excluded due to systematic output failures and are discussed in [Section 6.3](#).

As a supervised baseline, we reproduce the BERT-based fine-tuned models from [Chen et al. \(2024a\)](#), using leave-one-provider-group-out cross-validation and

class-weighted loss. We varied the learning rate (1e-5 to 6e-5), loss parameter ($\gamma = \{0, 1, 2\}$) and weight decay ($\{0, 0.01, 0.1\}$). Across all the configurations, due to instability caused by extreme imbalance, signals with high variance (>0.2) across folds are treated as non-modelable.

Finally, we train a lightweight logistic regression ensemble over all LLM configurations, to test whether diverse model-prompt configurations yield complementary predictions.

4. Analyses

We evaluate language models as text-based sensors of social behavior by comparing their predictions against expert annotations across 20 social signals. Due to extreme class imbalance across nearly all signals, we report *balanced accuracy*, which equally weighs performance on majority and minority classes and is standard for behavioral sensing tasks with skewed label distributions such as the ones showed by Xu et al. (2024) and Kaufman et al. (2025).

4.1. Overall Performance Under Severe Imbalance

Across all model-prompt configurations, social signal prediction is fundamentally constrained by data imbalance making fine-tuning impractical for many signals. In contrast, prompted large language models achieve above-chance balanced accuracy for **all 20 social signals** without any task-specific training. While absolute performance varies across signals, these results indicate that LLMs encode prior knowledge about interpersonal dynamics that can be elicited through prompting. Many signals, particularly Type-II affective states such as irritation, nervousness, and emotional distress, are rare or nearly absent in the dataset. However LLMs perform strongly despite these imbalances. While the large LLaMA with few-shot prompting provides best performance over all the subtasks, there is **no clear one-configuration-fits-all** for all the subtasks. This is further highlighted when the naive-ensemble betters most individual models.

4.2. LLM behavior across patient race

Since social signals are closely linked to patient race, often due to providers' implicit biases as shown by Cooper et al. (2012) and Bedmutha et al. (2023b),

we examine whether the LLM performance varies between White and non-White patients. This formulation allows us to frame this evaluation as a fairness investigation.

(1) Performance Gap between White and non-White patients: Our dataset contains 74 visits involving White patients and 17 involving non-White patients. Despite differences in label distributions between the groups, we observe that most balanced accuracy scores (computed over all the configurations) across both groups remain above 0.5 but show substantial predictive disparity. Only five social signals scored below 0.5, indicating that in general LLMs tend to be able to track social signals for both demographic subgroups. However, certain signals were entirely absent in the non-White group, which likely contributes to the larger observed performance gaps for example, *provider hurriedness* ($\bar{\Delta} = 0.235$) and *provider irritation* ($\bar{\Delta} = 0.130$). Interestingly, even when the label distributions were similar, we found that LLMs were better able to detect patient empathy among non-White patients ($\bar{\Delta} = -0.110$), suggesting possible unintended biases in model attention or interpretation. Table E1 in the Appendix shows complete results.

(2) Perceived Fairness of Individual Model Predictions: We investigate the fairness of the model predictions through Demographic Parity Ratio between White and non-White patients. We apply the four-fifths rule as a threshold, which indicates disparate impact occurring if the performance ratio between the groups falls below 4/5. The parity ratios are presented in Figure 2. We find very few instances of disparity: FLAN-T5 and Gemma configurations show no evidence of disparate impact, while Llama exhibits disparities on six social signals—specifically, *provider dominance*, *patient dominance*, *patient interactivity*, *provider hurriedness*, *provider irritation*, and *patient irritation*. These results suggest that Llama may be sensitive to linguistic cues that correlate with patient demographics which are often linked to broader social determinants of health.

4.3. Does performance shift across stages of a medical visit?

Clinic visits usually follow a predictable flow. Robinson (2003) described this as a series of phases: opening (greetings, patient concerns), middle (detailed symptom discussion and diagnosis), and clos-

Social Signal	BERT (fine-tuned)	FLAN-T5 ZS	FLAN-T5 FS	Gemma ZS	Gemma FS	Gemma COT	LLaMA ZS	LLaMA FS	LLaMA COT	LLaMA FSCOT	Ensemble (LOGO)
<i>Type-I Signals</i>											
Provider Dominance	0.618	<u>0.549</u>	0.539	0.522	0.502	0.478	0.507	0.498	0.538	0.495	0.603 <small>(0.13)</small>
Provider Attentiveness	0.563	0.531	0.520	0.509	0.489	0.523	0.463	<u>0.536</u>	0.437	0.523	0.624 <small>(0.09)</small>
Provider Warmth	0.571	0.467	0.469	0.518	0.481	0.506	0.517	<u>0.545</u>	0.512	0.511	0.558 <small>(0.10)</small>
Provider Engagement	0.543	0.489	0.489	0.459	0.483	0.500	0.513	<u>0.556</u>	0.534	0.564	0.609 <small>(0.19)</small>
Provider Empathy	0.500	0.530	0.532	0.516	0.579	0.489	0.610	<u>0.623</u>	0.585	0.629	0.628 <small>(0.13)</small>
Provider Respect	N/A	0.494	0.499	0.509	0.532	0.552	<u>0.542</u>	0.512	0.529	0.495	0.488 <small>(0.06)</small>
Provider Interactivity	0.643	0.510	0.519	0.499	0.512	0.512	0.530	0.539	<u>0.560</u>	0.542	0.484 <small>(0.06)</small>
Patient Dominance	0.583	0.524	0.516	0.471	0.491	0.500	<u>0.601</u>	0.554	0.608	0.568	0.587 <small>(0.08)</small>
Patient Attentiveness	0.549	0.520	0.518	0.496	0.476	0.536	0.516	0.577	0.515	<u>0.576</u>	0.589 <small>(0.03)</small>
Patient Warmth	0.539	0.473	0.477	0.488	0.510	0.505	<u>0.577</u>	0.557	0.547	0.582	0.564 <small>(0.06)</small>
Patient Engagement	0.614	0.549	0.548	0.462	0.513	0.480	0.553	<u>0.599</u>	0.557	0.595	0.554 <small>(0.04)</small>
Patient Empathy	N/A	0.566	0.570	0.393	0.483	0.491	0.456	0.487	0.589	<u>0.575</u>	0.464 <small>(0.18)</small>
Patient Respect	0.599	0.490	0.499	0.514	0.413	0.529	0.496	<u>0.546</u>	0.513	0.495	0.575 <small>(0.10)</small>
Patient Interactivity	0.705	0.540	<u>0.544</u>	0.454	0.442	0.515	0.492	0.488	0.500	0.480	0.599 <small>(0.13)</small>
Provider Hurriedness	N/A	0.350	0.358	<u>0.555</u>	0.446	0.568	0.551	0.492	0.267	0.448	0.664 <small>(0.24)</small>
<i>Type-II Signals</i>											
Provider Irritation	N/A	0.585	0.457	0.550	0.125	0.562	0.857	0.947	<u>0.927</u>	0.734	0.901 <small>(0.18)</small>
Patient Irritation	N/A	0.520	0.527	0.547	0.465	0.362	0.685	<u>0.631</u>	0.583	0.612	0.659 <small>(0.19)</small>
Patient Nervousness	0.600	0.492	0.498	0.404	0.489	0.379	0.514	0.726	0.586	<u>0.608</u>	0.650 <small>(0.23)</small>
Patient Sadness	0.600	0.545	0.550	0.516	0.469	0.461	0.624	0.686	<u>0.696</u>	0.711	0.681 <small>(0.15)</small>
Patient Distress	0.500	0.525	0.515	0.508	0.549	0.441	0.568	0.659	<u>0.639</u>	0.633	0.644 <small>(0.10)</small>
MEAN	<u>0.581</u>	0.512	0.507	0.494	0.473	0.494	0.559	0.588	0.561	0.569	0.606
STD	0.05	0.05	0.04	0.04	0.09	0.05	0.09	0.11	0.12	0.07	0.09

Table 1: Table shows the overall performance of different models and configurations as balanced accuracies. We use fine-tuned BERT (Chen et al., 2024a) for comparison with LLM predictions. We also train an ensemble model based on the predictions of LLMs over groups of providers and report additive effects in performance. Values in **bold** are the best-performing model and underline refers to the second-best model. In binary settings, a balanced accuracy greater than 0.5 implies the model performs better than chance

ing (summarization and treatment planning). Since each segment has a set pattern in conversations, the LLM performance may vary and be indicative of the underlying content. We divide each visit into **three segments: the start (first three-minute slice), middle (all intermediate slices), and end (last slice)**. Our subsequent analysis explores this variation in manually-coded behaviors and model predictions:

We examine whether the distribution of binary social signal labels differs across three *segments* of each visit: start, middle, and end. For each segment, we compute balanced accuracy across all configurations (Figure 3). Language models perform **least accurately at the start**, achieve **highest performance in the middle**, and show a modest decline toward the end; likely because the middle involves more balanced interactive exchanges that provide richer contextual cues for LLM-based social behavior inferences.

5. Post-Hoc Ensembling by Contextual Aggregation

Section 4 showed that LLMs demonstrate strong inherent capability for inferring social signals from clinical transcripts, often matching or exceeding fine-tuned task-specific models. However, performance varies substantially across signals, patient race, and temporal segments of a visit, and no single configuration consistently dominates. Therefore real-world deployment will be limited by task-specific and context-dependent gaps in performance. This motivates ensemble-based inference across multiple model-prompt configurations to attain robustness.

In clinical deployments, models are often accessed via APIs exposing only discrete predictions without probabilities. Most approaches of uncertainty or confidence estimation rely on underlying distribution (Huang et al., 2024) therefore soft-voting is infeasible. We therefore propose an agreement-weighted ensemble that operates under query-only access and leverages auxiliary metadata available at inference time.

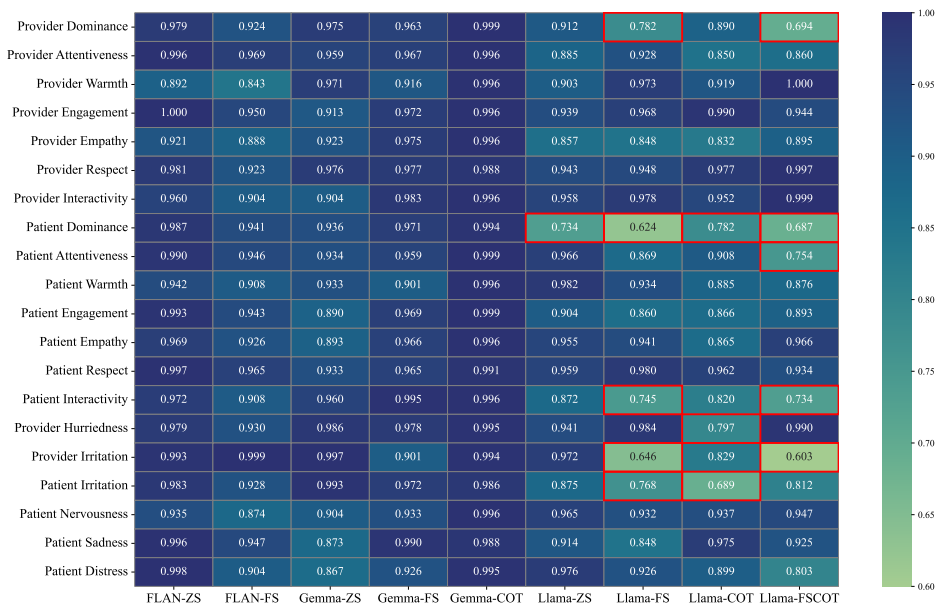


Figure 2: Demographic Parity Ratio between white (n=74) and non-white (n=17) patients. Most configurations follow the four-fifths rule (threshold = 0.8), but some disparity was observed for LLAMA.

5.1. Assumptions

We make these assumptions about our setting.

- 1. Query-access to configurations:** Multiple models and prompts can be queried, but only discrete predictions are returned.
- 2. Auxiliary information:** Patient race and transcript position within the visit (start, middle, end) are available at inference time.
- 3. Benchmark statistics:** A human-annotated benchmark provides performance estimates for each configuration (Table 1).
- 4. Dataset imbalance:** Label rarity and demographic imbalance may influence reliability of performance estimates.

5.2. Group-Level Agreement Estimation

From Section 4, it is evident that benchmark performance for a task or subgroup (such as race or time-of-visit) can be misleading due to disparities in the dataset. Therefore, we will have less confidence in a configuration that performs suboptimally or a group that is less represented in the benchmark. Thus for each group in the known dataset, if we can weigh individual samples by their perceived extent of **agreement with ground truth**, the ensemble will aim to achieve the best performance while being aware of

the current capabilities of the model-prompt configurations available. Note, in subsequent description, we refer to each configuration as an individual *model*.

Conventional estimates of agreement, uncertainty, or confidence typically rely on probability outputs, as reviewed by Huang et al. (2024). However, Chen and Mueller (2023) showed that uncertainty can alternatively be estimated by evaluating consistency across prompts or model variations. Inspired by this perspective, we estimate net agreement within groups conditioned on available auxiliary information.

Let x_0 denote transcript input, y the true label, \hat{y} the model prediction, and x_1, \dots, x_n auxiliary variables defining groups. Agreement is defined as:

$$\Pr(\text{agree} \mid x_0, \dots, x_n; \theta) = \sum_{k=0}^K \Pr(\hat{y} = k \cap y = k \mid x_0, \dots, x_n; \theta) \quad (1)$$

Using the chain rule:

$$\Pr(\text{agree} \mid \mathbf{x}_{0:n}; \theta) = \sum_{k=0}^K \underbrace{\Pr(\hat{y} = k \mid y = k, \mathbf{x}_{1:n}; \theta)}_{\text{model precision for group}} \times \underbrace{\Pr(y = k \mid \mathbf{x}_{1:n})}_{\text{label prevalence}} \times \underbrace{\Pr(\mathbf{x}_{1:n} \mid x_0)}_{\text{group prevalence}} \quad (2)$$

The first term reflects model precision conditioned on group membership, while the remaining terms

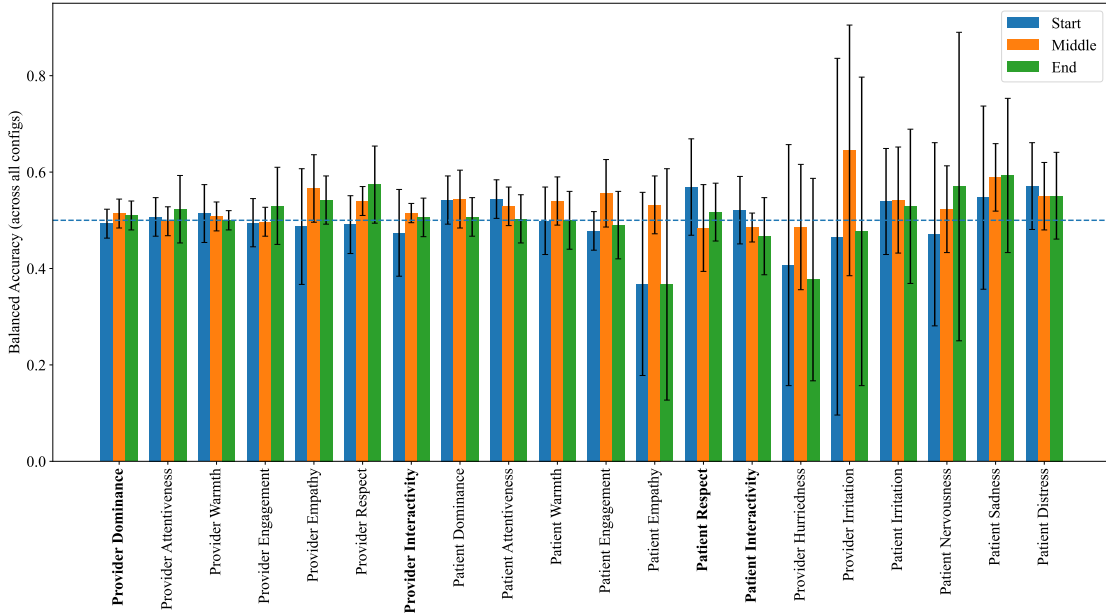


Figure 3: Figure shows the variability in performance at different segments (start, middle, end) of a visit. The bars represent average performance over all the model configurations and standard deviations describe the variability within the models. Signals marked in **bold** describe the ones whose distribution in a visit varies significantly over the segments

capture label imbalance and subgroup prevalence. Agreement therefore accounts for both task imbalance and demographic disparities.

5.3. Agreement-Weighted Ensemble Construction

We consider binary classification ($K \in \{-1, 1\}$). Auxiliary variables include race ($X_1 \in \{\text{White, Non-White}\}$) and visit segment ($X_2 \in \{\text{Start, Middle, End}\}$), yielding six mutually exclusive groups.

For each training sample l , model θ_m outputs $\theta_m(x_0^l) \in \{-1, 1\}$. Agreement weights $w_{g,m}$ are estimated per model and group using Eq. 2. Each sample inherits its group weight ($w_{l,m} = w_{g,m}$), and scaled predictions are computed as $s_{l,m} = w_{l,m} \cdot \theta_m^l$. An aggregator function f is trained on scaled scores across models. Algorithm 1 describes the approach stepwise.

The inference is similarly driven, by first assigning a slice to a group and then invoking the f for that setting.

Experimental Setup We evaluate using leave-one-group-out cross-validation across five provider groups to avoid leakage. Aggregators include Logistic Regression and Decision Trees (max depth=2).

Algorithm 1 Ensemble Training for each task

Input: Training data $\{(x_0^l, x_1^l, x_2^l, y^l)\}_{l=1}^N$, models $\{\theta_m\}_{m=1}^M$
Output: Group-level weights $w_{g,m}$
 Define groups $\mathcal{G} = \{\text{White, Non-White}\} \times \{\text{Start, Middle, End}\}$
foreach $g \in \mathcal{G}$ **do**
 for $m \leftarrow 1$ **to** M **do**
 Collect predictions $\theta_m(x_0^l)$ for samples with $(x_1^l, x_2^l) \in g$
 Compute agreement $w_{g,m}$ using Eq. 2
 end
end
 Fit aggregator f on $\{s_{l,m} = w_{g(l),m} \theta_m(x_0^l)\}$ **return** $\{w_{g,m}\}$ and f

We vary top- k models ($k = 2, 3, 5, 9$ [all]) ranked by benchmark performance. A naive ensemble baseline uses identical weights ($w = 1$).

5.4. Ensembled SSP Performance Evaluation

We describe the effect of aggregation using three approaches for all the 20 signals: (i) Raw or individual best performing model, (ii) LOGO, a leave-one-

group-out ensemble baseline using logistic regression without any group-level adjustments (i.e. $w = 1$ for all inferences), and (iii) Ensemble (Best) representing the best-performing configuration of agreement-weighted ensembles. The raw scores describe the idea of choosing the only best known model in the benchmark for the task. LOGO ensembles over all the models but unaware of the group-level (auxiliary) information and the Ensemble scores show the effect of group-level aggregation.

Figure 4 shows balanced accuracy across all social signals for the three approaches. Averaged over 20 signals and folds, the mean balanced accuracy for Raw (Best) is 0.614 (SD = 0.063), for LOGO is 0.606 (SD = 0.124), and for Ensemble (Best) is 0.664 (SD = 0.063).

Aggregation is inherently beneficial. On comparing the RAW and LOGO we observe that even even simple aggregation enhances performance for most signals. The LOGO ensemble matches or exceeds the best individual model, with particularly notable gains for *patient attentiveness*, *patient dominance*, and *provider hurtfulness*, where ensemble averaging clearly stabilizes and improves predictions.

Group-weighted aggregation provides substantial improvements. With the use of added contextual information, the ENSEMBLE approach a mean balanced accuracy of 0.664 an 5% increase over the best single model (Llama-FS) and a 5.8% improvement compared to LOGO. In Figure 4, in fact we note that in 16 out of 20 signals, the group-weighted aggregation results in the best performing model. This improvement is achieved by maintaining comparable if not lower deviation over folds.

These results demonstrate that while aggregation is inherently valuable, *how* we aggregate matters significantly. By leveraging contextual information (race, time-within-visit), the weighted ensemble selectively amplifies model strengths while maintaining stability—achieving best performance on 16 of 20 signals without increased variance and in fact reduces variance in 10 among them. This makes agreement-weighted ensembling particularly suitable for clinical deployment where reliability across diverse patient populations is essential.

6. Discussion

In this work, we introduced social signal processing (SSP) for clinical conversations using LLMs, demon-

strating feasibility across 20 signals and proposing contextual ensembling to address performance variability. Since the best performing configuration is very different for each task, we also note that SSP deployment is a benchmark-sensitive space. We discuss implications of our findings below.

6.1. LLM Orchestration for Clinical SSP

We know that every model tends to have varying degrees of performance and precision for each task. From our experiments in aggregation (Section 5), we observed that even within the three pipelines described, the best performing individual models and parameters of aggregation may vary based on factors such as the task, group (race, time-within-visit), number of models to be included (top-k), and aggregator (logistic regression or decision tree).

Additionally, on-device model may fail, or cloud models may experience downtime; an intelligent system could delay feedback for reliability, or provide immediate but less reliable feedback depending on context. Since we now have extensive information of the best performing model for each user, task and model, feedback systems can leverage task-specific insights on the ideal choice of parameters, base-models and prompts to personalize how to best address the users’ needs. Therefore, our comparative benchmarking (Table 1) and aggregation experiments (Appendix G, Table G1), can help interface designers route feedback through the ideal configurations, in such a way to meaningfully aggregate their results; ultimately this will allow them to design an optimal orchestration layer of LLM systems for SSP tasks.

6.2. Toward Trustworthy LLM Feedback

Understanding when the model is likely right (and when it may be wrong) is essential. When LLM outputs are shown as communication feedback (Steenstra et al., 2025), or used to score communicative behavior (La et al., 2025), designers need a sense of reliability. As Fargen et al. (2012) noted, unchecked alerts can cause notification fatigue, especially for clinicians. Park et al. (2023) showed that selective confidence-based feedback can also improve trust and adoption.

While group-agreement scores (Equation 2) provide a coarse confidence measure, linguistic cues could enable finer-grained “difficulty scores” for more nuanced confidence estimation.

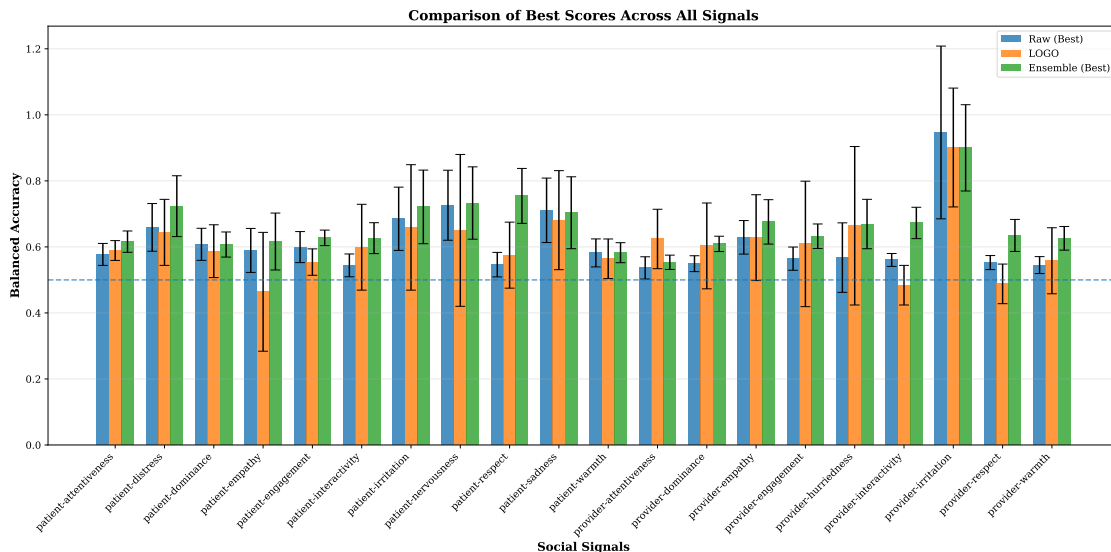


Figure 4: Comparison of balanced accuracy across all 20 social signals for three methods: Raw (Best) shows the best-performing individual model per task, LOGO represents a leave-one-group-out logistic regression ensemble without group-level weighting, and Ensemble (Best) is our agreement-weighted ensemble approach. Error bars indicate standard deviation.

Such a metric could help inform whether feedback should be pushed to a user, delayed, or adapted ultimately supporting safer and more trustworthy integration of LLM-based feedback systems into high-stakes environments like clinical care.

6.3. Limitations and Future Work

While our work demonstrates the promise of LLMs in modeling social behaviors in clinical conversations, several important limitations must be acknowledged. These constraints highlight opportunities for further research and refinement, and point to the necessary steps toward deploying social signal processing (SSP) systems that are robust, fair, and clinically useful.

First, we conducted our experiments on an annotated subset of the *Establishing Focus (EF)* dataset, which, although valuable, was collected nearly two decades ago. The dataset lacks a formal, codified annotation guide for its 21 social signals, relying instead on a *trainer-trains-trainer* method where new annotators learned from predecessors until sufficient agreement was reached. This introduces potential label drift and limits our ability to design prompt strategies around standardized definitions—something more feasible in aligned domains such as motivational interviewing (Steenstra et al., 2025; Shah et al., 2022) or depression detection (Chen et al.,

2024b). Nevertheless, the EF dataset remains one of the few with real-world, RIAS-style multi-signal annotations, making it a valuable starting point.

A more critical limitation stems from the high class imbalance across most social signals. Several affective states—such as *provider nervousness*—were either completely absent or extremely rare, forcing us to recast the task as a binary classification problem. This imbalance is likely exacerbated by the Hawthorne effect (Landsberger, 1958). Promising alternatives include data from standardized patients, as done by Bedmutha et al. (2024b), which we are exploring to elicit difficult conversations in realistic and controlled settings.

From a modeling perspective, we encountered generation failures in smaller models. For instance, FLAN-T5-Base failed to produce coherent responses in Chain-of-Thought prompting, and Gemma2-2B struggled in FS-CoT configurations. Although their raw logits could be extracted for classification, their outputs lacked full interpretability and consistency, making qualitative or explanation-based evaluation impractical. These limitations highlight the importance of model expressiveness—not just accuracy—in sensitive behavioral tasks.

In formative fine-tuning experiments, we found that larger models often overfitted to the majority class, especially when constrained by smaller batch

sizes on limited hardware. Despite efforts using dropout, LoRA (Hu et al. (2022)), and frozen encoders, fine-tuned performance did not surpass zero- or few-shot prompting in a meaningful way. This informed our decision to focus on maximizing performance via prompting, and to treat LLMs as *general-purpose social interpreters* rather than narrow classifiers. As more datasets emerge, our ongoing work will also include fine-tuning medium sized models.

7. Conclusion

Interpersonal communication is central to quality care, yet assessing social behavior at scale is difficult. We present a no-fine-tuning pipeline that extracts 20 affective social signals directly from clinical transcripts. Across three models and four prompting strategies, many configurations remained reliable despite severe label imbalance. Larger models benefited most from prompt augmentation, while smaller models matched or exceeded performance on specific tasks. Performance shifted across patient race and visit segment, with LLaMA showing disparities on six signals while FLAN-T5 and Gemma showed none.

To address variability under realistic constraints, we propose: under assumption of a query-access system, an agreement-weighted ensemble that uses demographic- and segment-specific agreement weights. This approach improved balanced accuracy by 5% over the best single model and 5.8% over an unweighted ensemble, outperforming the best individual model on 16/20 signals while requiring only binary outputs. Overall, this work demonstrates a practical method for reliable, transcript-based social signal modeling in clinical care.

8. Ethical Implications and Responsible Usage

Our work uses real-world clinical recordings annotated by expert raters; biases due to patient and score imbalances may impact findings. We maintained a HIPAA-compliant environment, and any deployment should ensure end-to-end data security. Finally, while sensing communication skills benefits providers, such tools could be misused for surveillance by administrators—data privacy for each inference is critical.

Acknowledgments

This study was a part of the UnBIASED project which was funded by the National Institutes of Health, National Library of Medicine, project R01LM013301. We thank the funding agency and all of our collaborators on the UnBIASED project, especially co-investigators Wanda Pratt, Janice Sabin and Brian Wood. Bedmutha was also supported by research fellowships from the Jacobs Center for Health Innovation and the Sanford Institute for Empathy and Compassion at UC San Diego.

References

- Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*, 2025.
- AHRQ. Effects of establishing focus in the medical interview (r01hs 013172 pi lynne robins), March 2006. URL <https://www.ahrq.gov/sites/default/files/2024-07/robins-report.pdf>. Accessed October 9, 2024.
- John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596, 2023.
- Emily Bascom, Reggie Casanova-Perez, Kelly Tobar, Manas Satish Bedmutha, Harshini Ramaswamy, Wanda Pratt, Janice Sabin, Brian Wood, Nadir Weibel, and Andrea Hartzler. Designing communication feedback systems to reduce healthcare providers’ implicit biases in patient encounters. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.
- Manas Satish Bedmutha, Poorva Satish Bedmutha, and Nadir Weibel. Privacy-aware respiratory symptom detection in-the-wild. In *Adjunct Proceedings of the 2023 ACM International Joint*

- Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, page 428–432, New York, NY, USA, 2023a. Association for Computing Machinery. ISBN 9798400702006. doi: 10.1145/3594739.3610733. URL <https://doi.org/10.1145/3594739.3610733>.
- Manas Satish Bedmutha, Amrit Bhat, Sabrina Mangal, Emily Bascom, Wanda Pratt, Brian Wood, Janice Sabin, Nadir Weibel, and Andrea Hartzler. Towards inferring implicit bias in clinical interactions using social signals. *AMIA Annual Symposium. AI Showcase Stage III*, 2023b.
- Manas Satish Bedmutha, Emily Bascom, Kimberly R Sladek, Kelly Tobar, Reggie Casanova-Perez, Alexandra Andreiu, Amrit Bhat, Sabrina Mangal, Brian R Wood, Janice Sabin, et al. Artificial intelligence-generated feedback on social signals in patient-provider communication: technical performance, feedback usability, and impact. *JAMIA open*, 7(4):ooae106, 2024a.
- Manas Satish Bedmutha, Anuujin Tsendenbal, Kelly Tobar, Sarah Borsotto, Kimberly R Sladek, Deepansha Singh, Reggie Casanova-Perez, Emily Bascom, Brian Wood, Janice Sabin, et al. Conversense: An automated approach to assess patient-provider interactions using social signals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2024b.
- Sudershan Boovaraghavan, Haozhe Zhou, Mayank Goel, and Yuvraj Agarwal. Kirigami: Lightweight speech filtering for privacy-preserving activity recognition using audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–28, 2024.
- Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10:1–47, 2022.
- Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 7124–7128. IEEE, 2020.
- Feng Chen, Manas Satish Bedmutha, Ray-Yuan Chung, Janice Sabin, Wanda Pratt, Brian R Wood, Nadir Weibel, Andrea L Hartzler, and Trevor Cohen. Toward automated detection of biased social signals from the content of clinical conversations. *arXiv preprint arXiv:2407.17477*, 2024a.
- Feng Chen, Dror Ben-Zeev, Gillian Sparks, Arya Kadakia, and Trevor Cohen. Detecting ptsd in clinical interviews: A comparative analysis of nlp methods and large language models. *arXiv preprint arXiv:2504.01216*, 2025.
- Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *arXiv preprint arXiv:2308.16175*, 2023.
- Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tiejun Qian, and Minlie Huang. Depression detection in clinical interviews with llm-empowered structural element graph. 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 8181–8194, 2024b.
- Lisa A Cooper, Debra L Roter, Kathryn A Carson, Mary Catherine Beach, Janice A Sabin, Anthony G Greenwald, and Thomas S Inui. The associations of clinicians’ implicit attitudes about race with medical visit communication and patient ratings of interpersonal care. *American journal of public health*, 102(5):979–987, 2012.
- Kyle M Fargen, Timothy O’Connor, Steven Raymond, Justin M Sporrer, and William A Friedman. An observational study of hospital paging practices and workflow interruption among on-call junior neurological surgery residents. *Journal of graduate medical education*, 4(4):467–471, 2012.
- Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gašić. Affect recognition in conversations using large language models. *arXiv preprint arXiv:2309.12881*, 2023.
- AL Hartzler, RA Patel, M Czerwinski, W Pratt, A Roseway, N Chandrasekaran, and A Back. Real-time feedback on nonverbal clinical communication. *Methods of information in medicine*, 53(05):389–405, 2014.

- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. A survey of uncertainty estimation in llms: Theory meets practice. *arXiv preprint arXiv:2410.15326*, 2024.
- Robert A Kaufman, Emi Lee, Manas Satish Bedmutha, David Kirsh, and Nadir Weibel. Predicting trust in autonomous vehicles: Modeling young adult psychosocial traits, risk-benefit attitudes, and driving factors with machine learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2025.
- Hannah La, Ziming Li, Ha-Kyung Kong, and Roshan L Peiris. Exploring the efficacy of a chatbot training application in alleviating graduate students’ public-speaking anxiety during q&a. 2025.
- Henry A Landsberger. Hawthorne revisited: Management and the worker, its critics, and developments in human relations in industry. 1958.
- Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*, 2023.
- Chunfeng Liu, Karen M Scott, Renee L Lim, Silas Taylor, and Rafael A Calvo. Eqclinic: a platform for learning communication skills in clinical consultations. *Medical education online*, 21(1):31801, 2016.
- Man Luo, Christopher J Warren, Lu Cheng, Haidar M Abdul-Muhsin, and Imon Banerjee. Assessing empathy in large language models with real-world physician-patient interactions. In *2024 IEEE International Conference on Big Data (Big-Data)*, pages 6510–6519. IEEE, 2024.
- Edwin C Montiel-Vázquez, Christian Arzate Cruz, Jorge Adolfo Ramírez Uresti, and Randy Gomez. Empatheticexchanges: Towards understanding the cues for empathy in dyadic conversations. *IEEE Access*, 2024.
- Junghwan Park, Meelim Kim, Mohamed El Mistiri, Rachael Kha, Sarasij Banerjee, Lisa Gotzian, Guillaume Chevance, Daniel E Rivera, Predrag Klasnja, Eric Hekler, et al. Advancing understanding of just-in-time states for supporting physical activity (project justwalk jitai): protocol for a system id study of just-in-time adaptive interventions. *JMIR Research Protocols*, 12(1):e52161, 2023.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *OpenAI*, 2022. URL <https://openai.com/research/whisper>.
- Jeffrey D Robinson. An interactional structure of medical activities during acute visits and its implications for patients’ participation. *Health communication*, 15(1):27–59, 2003.
- Debra Roter and Susan Larson. The roter interaction analysis system (rias): utility and flexibility for analysis of medical interactions. *Patient education and counseling*, 46(4):243–251, 2002.
- Debra L Roter, Judith A Hall, Danielle Blanch-Hartigan, Susan Larson, and Richard M Frankel. Slicing it thin: new methods for brief sampling analysis using rias-coded medical dialogue. *Patient education and counseling*, 82(3):410–419, 2011.
- Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24, 2022.
- Ian Steenstra, Farnaz Nouraei, and Timothy W Bickmore. Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations. *arXiv preprint arXiv:2502.18673*, 2025.
- Richard L Street Jr, Howard Gordon, and Paul Haidet. Physicians’ communication and perceptions of patients: is it how they look, how they talk, or is it just the doctor? *Social science & medicine*, 65(3):586–598, 2007.
- G. Swain. Patients may suffer from hallucinations of ai medical transcription tools. *CIO*, 2024. URL <https://www.cio.com/article/3593403/patients-may-suffer-from-hallucinations-of-ai-medical-transcription.html>.

- Aaron A Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Balleca, Patricia Kipnis, Vincent Liu, and Kristine Lee. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, 5(3):CAT-23, 2024.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9, 2025.
- Alexandria Vail, Jeffrey Girard, Lauren Bylsma, Jeffrey Cohn, Jay Fournier, Holly Swartz, and Louis-Philippe Morency. Toward causal understanding of therapist-client relationships: A study of language modality and social entrainment. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 487–494, 2022.
- Steven D Vannoy and Lynne S Robins. Suicide-related discussions with depressed primary care patients in the usa: gender and quality gaps. a mixed methods analysis. *BMJ open*, 1(2):e000198, 2011.
- Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.
- Aditya B Vishwanath, Vijay Kumar Srinivasalu, and Narayana Subramaniam. Role of large language models in improving provider–patient experience and interaction efficiency: A scoping review. *Artificial Intelligence in Health*, page 4808, 2024.
- Zhiyuan Wang, Nusayer Hassan, Virginia LeBaron, Tabor Flickinger, David Ling, James Edwards, Congyu Wu, Mehdi Boukhechba, and Laura E Barnes. Commsense: A wearable sensing computational framework for evaluating patient-clinician interactions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–31, 2024.
- Rand R. Wilcox. The bootstrap. In *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*, pages 87–108. Springer, New York, NY, 2 edition, 2010.
- Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32, 2024.
- Haoning Xue, Wang Liao, and Jingwen Zhang. Interaction dynamics of social support expressions predict future support-seeking behaviors in online support groups. *Computers in Human Behavior*, 156:108224, 2024.
- Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, and Sophia Ananiadou. Mentallama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*, 2023.
- Andrew Zolensky, Kuk Jin Jang, Janice Sabin, Andrea Hartzler, Basam Alasaly, Sriharsha Mopidevi, Mark Liberman, and Kevin Johnson. Speaker role identification in clinical conversations. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 144–157. World Scientific, 2025.

Appendix A. Prompt Creation

We describe below the prompt settings used across different conditions and models. While the core task remains the same, we modified the output conditioning to reduce any failures on specific models.

- **Role Specification:** You are a behavior analyst assessing doctor patient interaction to help doctors with communication feedback. You will be given a small part of a transcript of a conversation between a doctor and a patient.
- **Job Description:** You have to score the <signal_name> in this text. Look at each behavior and look for deviations from normal or neutral behavior, be it positive (high) or negative (low).
- **Scoring Instructions:** [FLAN-T5/Gemma2] Is the `signal_name` higher than normal? [Llama] The expected scale tags each behavior with an integer: 0 means below average or not existent, 1 means above average or existent.
This is updated to "Did you see any presence of `signal_name` in this slice?" for Type-II signals.
- **Output Format (answer):** [Llama] Return the integer only. [FLAN-T5/Gemma2] Respond only with one word, "yes" or "no".
- **Output Format (answer + reasoning):** [Llama] Return the integer first, followed by an explanation of the score in two sentences. [FLAN-T5/Gemma2] Is the `signal_name` higher than normal? Start your answer with either Yes or No. Explain why you think it was higher or lower. DO NOT repeat any sentence!
- **Few-Shot Examples:**
 - *Llama:* Here are examples (please interpret speaker turns accurately based on the context, even if the diarization may not be perfect): High provider dominance/assertiveness example: xxx; Low provider dominance/assertiveness example: xxx.
 - *FLAN-T5/Gemma2:* #TRANSCRIPT: {example}; #LABEL: {yes}; ... #TRANSCRIPT: {example}; #LABEL: {no}; #TRANSCRIPT: {example}; #LABEL:

Appendix B. Robustness Investigation

Since the dataset is small and imbalanced, in addition to the individual results (Table 1), we conduct reliability estimation by *bootstrapping*. As prescribed in Wilcox (2010), we used 599 iterations to resample prediction and label pairs from each configuration and task. Figure 5 shows the results of this analysis.

For all the social signals at least one of the configurations result in an averaged balanced accuracy over all resamples greater than chance (0.5) with over 65.55% of all configuration-task pairs reporting above chance behavior. We also note that in 15 of 20 tasks, the lower bound of the 95% confidence interval is also greater than chance. The macro-F1 scores are a stricter metric and tend to be low in many signals. Overall these results further validate the **feasibility of tracking social signal levels** using LLMs.

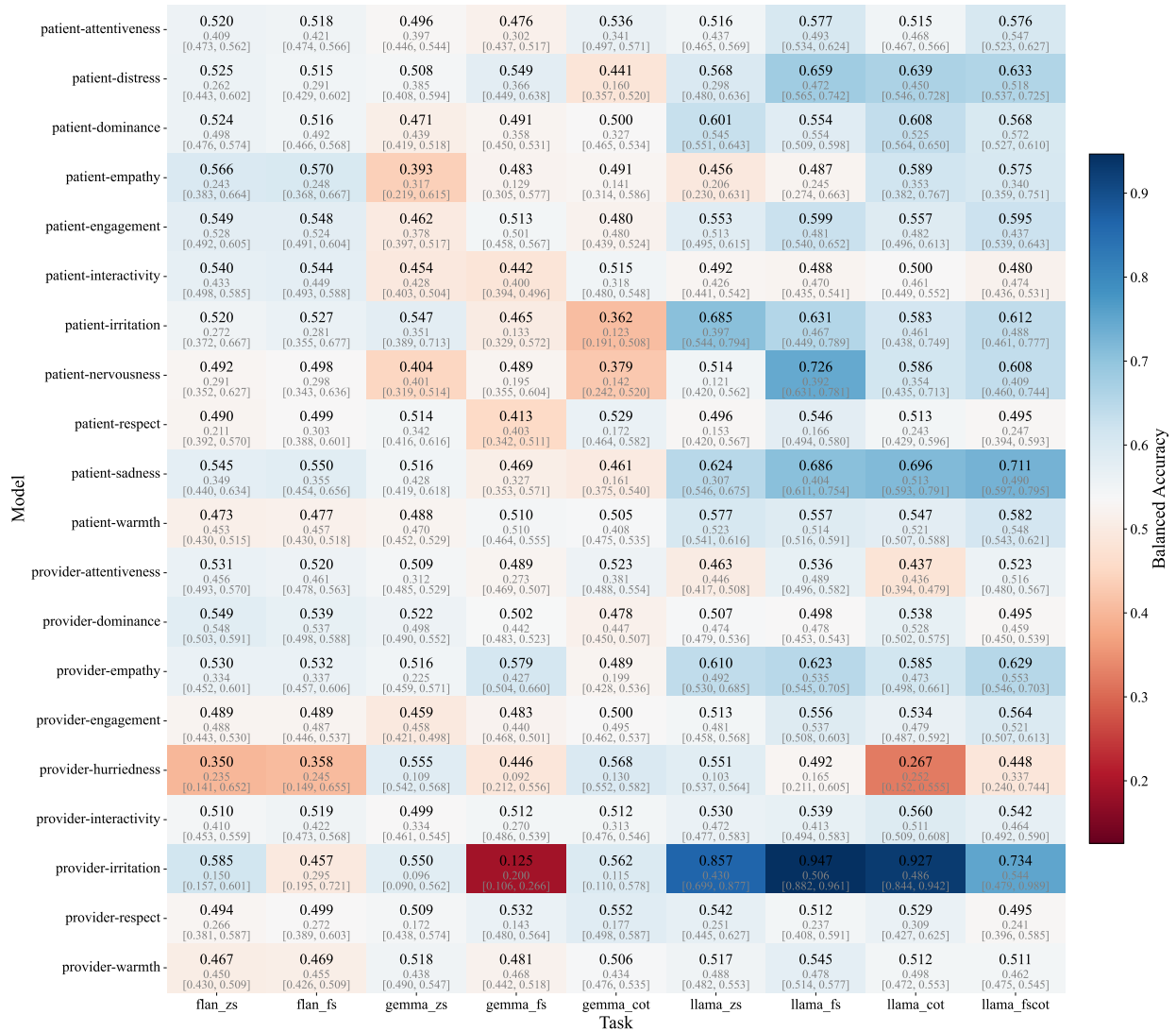


Figure 5: Bootstrapped performance for all configuration-task pairs is visualized in this heatmap, with each cell showing mean balanced accuracy, macro-F1, and 95% confidence intervals (based on balanced accuracy). At least one model per task reliably exceeds chance (0.5), and for most tasks, even the lower confidence bound is above 0.5, highlighting consistent predictive ability. Macro-F1 is provided for completeness. These findings demonstrate that LLMs can feasibly detect trends in social signals across tasks.

Appendix C. Indicators of a difficult slice

We also examine whether certain transcript segments—specific moments within a visit—are more difficult for LLMs when predicting social signals. Understanding which slices are consistently harder to interpret can help clarify when model outputs may be less reliable. These insights are useful for designing systems that surface uncertainty when providing feedback or assessments based on conversational data.

We first aggregate the number of correct predictions across all tasks to the level of individual transcript slices. With 9 model–prompt configurations and 20 tasks, each sample has a maximum of 180 possible predictions. The distribution of correct predictions per sample is approximately normal, as confirmed by a Kolmogorov–Smirnov test ($D = 0.048$, $p = 0.181$), with a mean of 81.39 and a standard deviation of 15.52. The number of correct predictions per sample ranges from 25 to 128. Figure 6 illustrates this distribution across all transcript slices. To better understand what makes certain samples easier or harder to interpret, we conduct two complementary analyses: one quantitative and one qualitative.

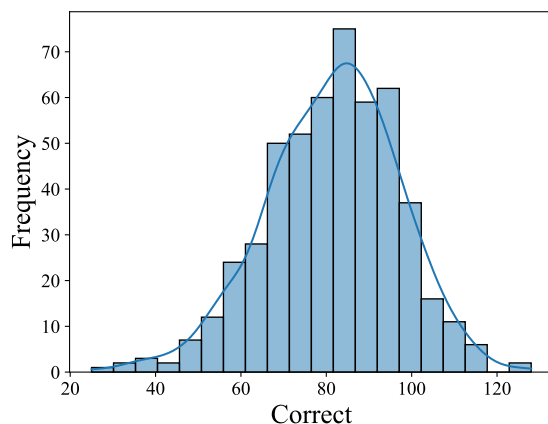


Figure 6: Histogram distribution of number of correct predictions per sample. We see that the correct predictions are normally distributed.

C.1. Quantitative Differentiation

Since our analysis is based entirely on textual transcripts, we examine linguistic characteristics to identify differences between high- and low-accuracy samples. To do this, we use LIWC-22 [Boyd et al. \(2022\)](#), a widely used tool for capturing affective and psychological dimensions of language. LIWC has been validated in related domains such as therapist–patient interactions [Vail et al. \(2022\)](#) and support-seeking conversations [Xue et al. \(2024\)](#).

We define “easy” and “hard” samples as slices falling in the upper and lower quantiles of the distribution of correct predictions, respectively. We then compare linguistic markers between these two groups using the Mann–Whitney U-test. To account for multiple comparisons across LIWC categories, we apply a Bonferroni correction. Table C1 presents the LIWC categories that significantly differ between “hard” and “easy” slices based on their agreement across all configurations and tasks. We observe that linguistic, cognitive, affective, and social-behavioral descriptors vary meaningfully between these groups. Interestingly, descriptors related to physical and mental health or lifestyle do not show significant differences. This suggests that the difficulty of social signal prediction is influenced less by the clinical content of the conversation and more by general linguistic and affective patterns. This highlights the potential for these findings to generalize across domains beyond healthcare.

Among linguistic features, more difficult slices tend to contain shorter sentences and exhibit higher usage of function words, impersonal pronouns, determiners, articles, and conjunctions, alongside lower frequencies

Feature	Description (example words)	Hard	Easy	U-statistic
Tone	Degree of positive (negative) tone	43.557 (26.14)	54.512 (28.86)	16956.5 *
WPS	Words per sentence	8.378 (4.03)	11.048 (47.32)	28704.5 **
function	Function words (the, to, and, I)	51.483 (9.87)	47.424 (10.72)	28676.5 **
you	Second-person pronouns (you)	3.994 (2.10)	6.684 (7.31)	17619.0 *
ipron	Impersonal pronouns (that, it, this, what)	6.845 (2.60)	5.718 (2.85)	28618.5 **
det	Determiners (the, at, that, my)	11.592 (2.91)	9.974 (4.06)	28189.0 *
article	Articles (a, an, the, alot)	4.335 (1.67)	3.624 (1.84)	28048.0 *
conj	Conjunctions (and, but, so, as)	6.208 (2.33)	5.231 (2.65)	28062.0 *
Cognition	Cognition (is, was, but, are)	12.167 (4.07)	10.377 (4.77)	27898.0 *
cogproc	Cognitive Processes (but, not, if, or, know)	10.546 (4.03)	8.747 (4.52)	28290.0 **
cause	Causation (how, becuase, make, why)	1.280 (0.88)	0.937 (0.73)	28615.5 **
tentat	Tentative (if, or, any, something)	2.843 (1.52)	2.205 (1.61)	28484.5 **
tone_pos	Positive Tone (good, well, new, love)	2.667 (3.65)	5.316 (7.87)	17581.0 *
polite	Politeness (thank(s), please, good morning)	0.607 (1.79)	3.545 (8.12)	16586.0 ***
AllPunc	Number of punctuation marks	32.920 (10.41)	35.459 (8.77)	17523.0 *
Period	Number of periods	13.686 (7.80)	15.979 (8.01)	17813.5 *

Table C1: Table shows the significantly different textual descriptors in the lower quantile (Hard) and higher quantile (Easy) of slices by their overall probability of correctness achieved by our system. The descriptions are the definitions and exemplar words from the LIWC-22 guide [Boyd et al. \(2022\)](#). We report the mean (std) for the LIWC feature values for the two groups as well as the Mann-Whitney U statistics with Bonferroni correction. We use the convention: * ($p < 0.05$), ** ($p < 0.01$) and *** ($p < 0.001$)

of second-person pronouns. In the cognitive domain, difficult samples show increased use of words associated with general cognitive processes (e.g., thinking, reasoning), causation, and tentativeness. In terms of affect, difficult samples are characterized by lower overall tone and reduced use of positive emotion words.

While punctuation-related features also differed between groups, we treat these results with caution, as they are likely influenced by transcription artifacts and may not generalize across datasets. One notable finding is that a lower presence of affective and tone-related words significantly reduces the LLM’s ability to detect social signals. This supports the interpretation that LLMs rely, at least in part, on affective linguistic cues when making predictions. The strong role of both cognitive and emotional language in distinguishing sample difficulty further suggests that LLMs are capable of abstracting away from the clinical topic itself and instead focus on general patterns of social behavior. Future work should explore how these trends differ by speaker role (e.g., patient vs. provider) to gain a more nuanced understanding of where and how LLMs succeed or fail in social signal processing.

C.2. Qualitative Understanding of difficult samples

To better understand the nature of moments that are challenging for LLMs to assess, we conducted a qualitative analysis of the lowest 10th percentile of samples ($n = 50$) by performance (see Figure 6). We explored two dimensions of these transcript slices: the content of the conversation and the style of communication. Using open coding, we inductively identified several recurring themes that characterize difficult-to-predict segments:

- *Anxious patients, reassuring providers:* Many difficult slices involve patients expressing anxiety—whether recalling experiences related to their health, voicing concerns about inherited conditions, or worrying about side effects or addiction to prescribed medication. In these cases, providers often adopt a reassuring role, attempting to reduce the patient’s distress by offering empathetic or explanatory responses. The subtlety and complexity of these emotional dynamics may make them harder for LLMs to interpret accurately.

- *Patient frustration with themselves or the healthcare System:* Another common theme is patient frustration, often stemming from feeling unheard or misunderstood by previous providers. For example, one patient recounted how multiple specialists attributed all symptoms to “arthritis,” while new symptoms went unacknowledged. In other cases, patients rapidly switched topics mid-sentence, disrupting the flow of the conversation and creating confusion. These disjointed or emotionally charged exchanges may present challenges for models trying to infer coherent social signals.
- *Actions in parallel to voice:* Several difficult slices involved segments in which providers appeared to be performing physical examinations or procedural tasks. These interactions consisted primarily of brief, instruction-oriented language with minimal dialogue or back-and-forth exchange. Such slices underscore a core limitation of using text-only inputs: without access to nonverbal context or action cues, LLMs struggle to interpret the social dynamics of these moments.

These themes reveal specific patterns in patient-provider communication that are difficult for LLMs to track using textual data alone. Recognizing these types of interactions may help inform the development of more adaptive systems—for example, by routing such segments through alternative multi-prompt strategies or incorporating multimodal signals to enhance robustness. Ultimately, this qualitative insight can guide future improvements in LLM-based social signal processing systems for healthcare and beyond.

Appendix D. Model Configurations’ impact on LLM performance

Our overall pipeline demonstrates the potential of LLMs to model social behavior. Here we investigate whether different model configurations yield different performance outcomes. This analysis helps us understand the relative impact of design choices and can guide future researchers in selecting optimal configurations for their specific applications.

Approach – We analyze the effects of four categorical variables: `model`, `prompt`, `configuration`, and `task`. Here, `task` refers to each of the 20 social signal prediction tasks. Also here the three models used are FLAN-T5, Gemma2-2B, and LLaMA3.1-405B, and the prompting strategies include Zero-Shot (ZS), Few-Shot (FS), Chain-of-Thought (CoT), and Few-Shot with Chain-of-Thought (FS-CoT). The `configurations` represent valid model–prompt pairings used during evaluation. To assess the relative contribution of each variable, we analyze the likelihood of a correct prediction (`correct`) as a function of these settings. A prediction is labeled `correct` if the model output matches the ground truth label for a given transcript-task pair. We then model the probability of correctness using Generalized Linear Mixed Models (GLMMs) with a binomial distribution. This method allows us to examine main effects while controlling for variability introduced by other factors. This novel framing of understanding alignment allows us to evaluate various factors impacting Human-AI alignment using known potential covariates. We use the `BinomialBayesMixedGLM` module from the Statsmodels library.¹ Equations 3a–3c show the model specifications for each variable under study, with the variable of interest treated as a fixed effect and all others treated as random effects.

This approach provides a more robust statistical framework than simply comparing accuracy metrics, which are often confounded by dataset imbalance or task difficulty. Prior work on LLMs in behavioral sensing Xu et al. (2024) has typically relied on such raw accuracy comparisons. By contrast, our analysis accounts for multi-dimensional variability and therefore helps more reliable inferences about the relative effectiveness of different model configurations.

1. Statsmodels documentation: [BinomialBayesMixedGLM](#)

$$\begin{aligned} \text{correct} &\sim C(\text{model}, \text{ref} = \text{“FLAN”}) - 1 \\ &\quad + \text{Random}(1 + \text{visit}, 1 + \text{prompt}, 1 + \text{task}) \end{aligned} \tag{3a}$$

$$\begin{aligned} \text{correct} &\sim C(\text{prompt}, \text{ref} = \text{“ZS”}) - 1 \\ &\quad + \text{Random}(1 + \text{visit}, 1 + \text{task}, 1 + \text{model}) \end{aligned} \tag{3b}$$

$$\begin{aligned} \text{correct} &\sim C(\text{configuration}, \text{ref} = \text{“FLAN-ZS”}) - 1 \\ &\quad + \text{Random}(1 + \text{visit}, 1 + \text{task}) \end{aligned} \tag{3c}$$

Effects of Model Settings — Table D1 presents the comparative effects of different model settings, including model types, prompting strategies, and their combined configurations. We use the smallest model (FLAN-T5) and the simplest prompting strategy (Zero-Shot) as reference baselines to assess the impact of varying each setting. Below, we summarize the key findings from each of these analyses.

	Coef. (log)	OR
<i>Model (from Eq 3a)</i>		
FLAN-T5 (reference)	-0.360	[0.698]
Gemma2-2b	-0.185	0.831
LLaMA3.1-405B	0.345	1.412
<i>Random Var (visit = 0.05, prompt = 0.22, task=0.25)</i>		
<i>Prompting Style (from Eq 3b)</i>		
Zero Shot (reference)	-0.356	[0.700]
Few Shot	0.094	1.098
Chain of Thought	-0.014	0.986
Few Shot with Chain of Thought	0.359	1.431
<i>Random Var (visit = 0.05, task=0.23, model = 0.12)</i>		
<i>Configuration (from Eq 3c)</i>		
FLAN-ZS (reference)	-0.199	[0.820]
FLAN-FS	-0.310	0.733
Gemma-ZS	-0.249	0.780
Gemma-FS	-0.401	0.670
Gemma-COT	-0.746	0.474
LLaMA-ZS	-0.178	0.837
LLaMA-FS	0.136	1.146
LLaMA-COT	0.252	1.286
LLaMA-FSCOT	0.402	1.495
<i>Random Var (visit = 0.05, task=0.23)</i>		

Table D1: Table shows the log-odds and Odds Ratios for Binomial Generalized Linear Mixed Models to predict the odds of being correct based on model-type, prompt-type and combinations of the two. The smallest model (FLAN-T5) and simplest prompt (ZS) are used as references. Odds ratios greater than 1 imply increase odds and vice versa.

1. *LLaMA is 1.4× more likely to be correct, while Gemma underperforms relative to FLAN-T5.* From Equation 3a in Table D1, we observe that, using FLAN-T5 as the reference, the odds ratio for Gemma2 is 0.831, whereas LLaMA achieves an odds ratio of 1.412. Random effects across visits are low, while moderate variability is observed across prompt styles and tasks. These results indicate that LLaMA,

across all prompting styles, consistently demonstrates better understanding of social signals in clinical conversations. In contrast, Gemma2 underperforms FLAN-T5 by approximately 17%, despite being a larger model. This suggests that FLAN-T5 may have a stronger inductive bias toward affective content (especially Type-I signals), but LLaMA is more reliable overall.

2. *Prompt-augmentation results are comparable to ZS.* According to Equation 3b in Table D1, FS-CoT achieves an odds ratio of 1.431, significantly outperforming the Zero-Shot (ZS) baseline. Few-Shot (FS) prompting shows a modest improvement (1.098), while Chain-of-Thought (CoT) slightly underperforms ZS (0.986). These results suggest that combining examples with reasoning (FS-CoT) yields the greatest benefit in social affective sensing tasks. However, this configuration requires models capable of producing coherent free-text justifications—something only LLaMA could achieve reliably in our setup. Randomization across models and visits has low impact, while variability across tasks is moderate. Together, these findings reinforce the value of combining in-context examples with reasoning to improve performance.

3. *Only LLaMA configurations with prompt augmentations outperform the FLAN-ZS baseline.* Using FLAN-ZS as the reference configuration, Equation 3c reveals that Gemma-based configurations consistently underperform, with odds ratios of 0.780 (Gemma-ZS), 0.670 (Gemma-FS), and 0.474 (Gemma-CoT). FLAN-FS also shows lower odds (0.733), while LLaMA configurations steadily improve with each augmentation: 0.837 (LLaMA-ZS), 1.146 (LLaMA-FS), 1.286 (LLaMA-CoT), and 1.496 (LLaMA-FSCoT). This configuration-level analysis—where each model-prompt pair is treated as a distinct unit—offers a more balanced comparison, as sample sizes are uniform across conditions. These results underscore that while FLAN-T5 outperforms several larger configurations (including Gemma and LLaMA-ZS), LLaMA benefits most from prompt augmentations and consistently achieves the highest performance when both few-shot examples and reasoning are included.

Appendix E. Variation over Race

Table E1 summarizes group-level differences in social signal distributions and model performance across visits involving White and non-White patients. The first set of columns reports the proportion of slices labeled as high for each signal, while the “Statistical Difference” column quantifies their relative divergence. Provider warmth and engagement exhibit the most notable distributional disparities, suggesting systematic variation in perceived affective behaviors across patient groups. The rightmost columns report balanced accuracies averaged over all configurations, illustrating that while performance is generally comparable across groups, some signals—such as provider empathy, patient empathy, and distress—show group-specific sensitivity. Bolded values (>0.55) indicate reliable above-chance tracking, whereas underlined values reflect cases where LLM-based inference underperforms a majority baseline.

Social Signal	Labels		Statistical Difference	Balanced Accuracy	
	White	Non-White		White	Non-White
Provider Dominance	0.666 (0.47)	0.761 (0.43)	1.597	0.505 (0.03)	0.557 (0.07)
Provider Attentiveness	0.332 (0.47)	0.326 (0.47)	0.975	0.501 (0.03)	0.514 (0.07)
Provider Warmth *	0.526 (0.50)	0.337 (0.48)	0.457	0.508 (0.03)	<u>0.481 (0.05)</u>
Provider Engagement **	0.808 (0.39)	0.630 (0.49)	0.406	0.509 (0.03)	<u>0.513 (0.06)</u>
Provider Empathy	0.079 (0.27)	0.087 (0.28)	1.105	0.584 (0.07)	<u>0.494 (0.06)</u>
Provider Respect	0.048 (0.21)	0.033 (0.18)	0.667	0.512 (0.03)	0.560 (0.14)
Provider Interactivity	0.212 (0.41)	0.293 (0.46)	1.548	0.531 (0.02)	0.502 (0.03)
Patient Dominance	0.293 (0.46)	0.152 (0.36)	0.433	0.530 (0.04)	0.571 (0.10)
Patient Attentiveness	0.267 (0.44)	0.130 (0.34)	0.412	0.528 (0.03)	0.500 (0.06)
Patient Warmth	0.454 (0.50)	0.293 (0.46)	0.499	0.523 (0.04)	0.521 (0.05)
Patient Engagement	0.846 (0.36)	0.859 (0.35)	1.105	0.547 (0.05)	0.505 (0.08)
Patient Empathy	0.012 (0.11)	0.011 (0.10)	0.903	<u>0.494 (0.05)</u>	0.604 (0.17)
Patient Respect	0.048 (0.21)	0.022 (0.15)	0.440	<u>0.502 (0.04)</u>	<u>0.473 (0.14)</u>
Patient Interactivity	0.219 (0.41)	0.293 (0.46)	1.484	0.502 (0.04)	<u>0.476 (0.05)</u>
Provider Hurriedness	0.012 (0.11)	0.000 (0.00)	0.000	0.450 (0.10)	<u>0.215 (0.15)</u>
Provider Irritation	0.005 (0.07)	0.000 (0.00)	0.000	0.637 (0.26)	<u>0.507 (0.36)</u>
Patient Irritation	0.022 (0.15)	0.011 (0.10)	0.497	0.542 (0.11)	0.589 (0.15)
Patient Nervousness	0.022 (0.15)	0.043 (0.21)	2.056	0.526 (0.16)	0.511 (0.09)
Patient Sadness	0.046 (0.21)	0.043 (0.21)	0.950	0.588 (0.11)	0.566 (0.10)
Patient Distress	0.043 (0.20)	0.109 (0.31)	2.696	0.584 (0.11)	0.515 (0.05)

Table E1: Table shows the differences in social signals across White and non-White patients. We see that when binarized, the social signal labels of provider warmth and engagement are significantly different. We also show the averaged balanced accuracies across all configurations on the two subset of visits reported as mean (std); **bolded values** show the best configurations (>0.55). Underlined Values are settings where LLM performance to track the social signal levels was worse than majority classifier (0.5)

Appendix F. Variation over Time within Visit

We divide each visit into three segments—start, middle, and end. The first and last slice of a visit are labeled as start and end, while all intermediate slices constitute the middle. We compare the distribution of human annotations and model performance as shown in Table F1.

Social Signal	Start (Dist)	Mid (Dist)	End (Dist)	Chi-square $\chi^2(2)$	Start (Acc)	Mid (Acc)	End (Acc)
Provider Dominance	0.527 (0.50)	0.752 (0.43)	0.582 (0.50)	■ 21.520 ***	<u>0.493 (0.03)</u>	0.514 (0.03)	0.510 (0.03)
Provider Attentiveness	0.363 (0.48)	0.324 (0.47)	0.330 (0.47)	† 0.477	<u>0.507 (0.04)</u>	<u>0.498 (0.03)</u>	0.523 (0.07)
Provider Warmth	0.714 (0.45)	0.418 (0.49)	0.549 (0.50)	■ 26.374 ***	0.514 (0.06)	<u>0.508 (0.03)</u>	0.500 (0.02)
Provider Engagement	0.769 (0.42)	0.791 (0.41)	0.714 (0.45)	† 2.401	<u>0.495 (0.05)</u>	<u>0.497 (0.03)</u>	0.530 (0.08)
Provider Empathy	0.033 (0.18)	0.100 (0.30)	0.055 (0.23)	■ 5.300	<u>0.487 (0.12)</u>	<u>0.566 (0.07)</u>	0.542 (0.05)
Provider Respect	0.099 (0.30)	0.030 (0.17)	0.055 (0.23)	■ 7.674	<u>0.491 (0.06)</u>	0.540 (0.03)	0.574 (0.08)
Provider Interactivity	0.055 (0.23)	0.303 (0.46)	0.110 (0.31)	■ 33.570 ***	<u>0.474 (0.09)</u>	0.515 (0.02)	0.506 (0.04)
Patient Dominance	0.253 (0.44)	0.276 (0.45)	0.275 (0.45)	† 0.197	<u>0.542 (0.05)</u>	0.544 (0.06)	0.507 (0.04)
Patient Attentiveness	0.209 (0.41)	0.245 (0.43)	0.253 (0.44)	† 0.620	0.544 (0.04)	0.529 (0.04)	0.503 (0.05)
Patient Warmth	0.495 (0.50)	0.394 (0.49)	0.484 (0.50)	■ 4.354	<u>0.499 (0.07)</u>	0.540 (0.05)	0.500 (0.06)
Patient Engagement	0.813 (0.39)	0.882 (0.32)	0.747 (0.44)	■ 10.795	<u>0.478 (0.04)</u>	0.556 (0.07)	<u>0.490 (0.07)</u>
Patient Empathy	0.000 (0.00)	0.015 (0.12)	0.011 (0.10)	–	<u>0.368 (0.19)</u>	0.532 (0.06)	<u>0.367 (0.24)</u>
Patient Respect	0.033 (0.18)	0.012 (0.11)	0.176 (0.38)	■ 44.923 ***	0.569 (0.10)	0.484 (0.09)	0.517 (0.06)
Patient Interactivity	0.088 (0.28)	0.303 (0.46)	0.110 (0.31)	■ 27.684 ***	0.521 (0.07)	<u>0.485 (0.03)</u>	0.467 (0.08)
Provider Hurriedness	0.011 (0.10)	0.009 (0.10)	0.011 (0.10)	† 0.044	<u>0.407 (0.25)</u>	<u>0.486 (0.13)</u>	<u>0.377 (0.21)</u>
Provider Irritation	0.000 (0.00)	0.006 (0.08)	0.000 (0.00)	–	<u>0.466 (0.37)</u>	<u>0.645 (0.26)</u>	<u>0.477 (0.32)</u>
Patient Irritation	0.033 (0.18)	0.018 (0.13)	0.022 (0.15)	† 0.743	<u>0.539 (0.11)</u>	0.542 (0.11)	<u>0.529 (0.16)</u>
Patient Nervousness	0.033 (0.18)	0.027 (0.16)	0.011 (0.10)	† 1.021	<u>0.471 (0.19)</u>	0.523 (0.09)	0.570 (0.32)
Patient Sadness	0.055 (0.23)	0.045 (0.21)	0.033 (0.18)	† 0.518	<u>0.547 (0.19)</u>	0.589 (0.07)	0.593 (0.16)
Patient Distress	0.066 (0.25)	0.058 (0.23)	0.033 (0.18)	† 1.106	<u>0.571 (0.09)</u>	0.550 (0.07)	0.551 (0.09)
Average (Acc)	–	–	–	–	0.499 (0.05)	0.532 (0.04)	0.507 (0.06)

Table F1: Merged summary of manually-coded binary label distributions and model accuracy across visit regions (start, middle, end). Distributions are shown as mean (SD). Significant Chi-squared tests with Bonferroni correction are marked (***) $p < 0.001$. Accuracy scores underlined fall below the 0.5 balanced accuracy threshold; bolded signals show significant distributional differences.

Table F1 shows the distribution of social behaviors across these segments and measures the extent to which they differ using a Chi-squared test with Bonferroni correction. Significant effects ($p < 0.001$) are observed for provider dominance, warmth, and interactivity, as well as for patients’ respect and interactivity, suggesting that interactional dynamics evolve notably across the course of the visit. The right-hand columns report the average balanced accuracy (and standard deviation) of model predictions within each segment. As shown, model performance remains consistent overall, with a slight improvement during the middle of the visit—potentially reflecting greater conversational richness or behavioral stability at that stage. Underlined values denote cases where balanced accuracy fails to exceed the 0.5 threshold.

Appendix G. Ensembling Experiments

Table G1 summarizes the best-performing configurations for each social signal across three evaluation modes: (1) *Raw* or best performing individual model; (2) *LOGO*, a leave-one-group-out logistic regression baseline without weighting; and (3) *Ensemble*, which aggregates predictions using agreement-weighted fusion of models with optimized classifier and selection strategies. Results are reported as balanced accuracy with standard deviation across held-out groups. As shown, ensemble methods consistently outperform both raw and LOGO baselines across most social signals demonstrating the benefit of leveraging complementary strengths of multiple prompting configurations and model families. Signals such as provider dominance, interactivity, and empathy—as well as patient distress and respect—show substantial gains under ensemble aggregation, highlighting their stability across model architectures and prompt variations.

Signal	Category	Method	Score	Std	Signal	Category	Method	Score	Std
provider-dominance	Raw	flan_zs	0.549	0.024	provider-hurriedness	Raw	gemma_cot	0.568	0.105
	LOGO	LOGO	0.603	0.130		LOGO	LOGO	0.664	0.240
	Ensemble	pr_wt_dtc_g	0.609	0.023		Ensemble	pr_wt_lr_l3p1	0.669	0.075
provider-attentiveness	Raw	llama_fs	0.536	0.034	provider-irritation	Raw	llama_fs	0.947	0.262
	LOGO	LOGO	0.624	0.090		LOGO	LOGO	0.901	0.180
	Ensemble	pr_mac_lr_l3p1	0.553	0.022		Ensemble	pr_wt_lr_all	0.900	0.131
provider-warmth	Raw	llama_fs	0.545	0.026	patient-irritation	Raw	llama_zs	0.685	0.096
	LOGO	LOGO	0.558	0.100		LOGO	LOGO	0.659	0.190
	Ensemble	pr_wt_dtc_l3p1	0.626	0.036		Ensemble	pr_wt_dtc_g	0.721	0.111
provider-engagement	Raw	llama_fscot	0.564	0.035	patient-nervousness	Raw	llama_fs	0.726	0.106
	LOGO	LOGO	0.609	0.190		LOGO	LOGO	0.650	0.230
	Ensemble	pr_wt_dtc_g	0.632	0.037		Ensemble	pr_wt_dtc_all	0.733	0.110
provider-empathy	Raw	llama_fscot	0.629	0.051	patient-sadness	Raw	llama_fscot	0.711	0.098
	LOGO	LOGO	0.628	0.130		LOGO	LOGO	0.681	0.150
	Ensemble	pr_wt_dtc_top3	0.676	0.067		Ensemble	pr_wt_lr_top2	0.703	0.109
provider-respect	Raw	gemma_cot	0.552	0.021	patient-distress	Raw	llama_fs	0.659	0.072
	LOGO	LOGO	0.488	0.060		LOGO	LOGO	0.644	0.100
	Ensemble	pr_mac_lr_top3	0.635	0.049		Ensemble	pr_wt_lr_top3	0.723	0.092
provider-interactivity	Raw	llama_cot	0.560	0.020	patient-dominance	Raw	llama_cot	0.608	0.049
	LOGO	LOGO	0.484	0.060		LOGO	LOGO	0.587	0.080
	Ensemble	pr_wt_dtc_top2	0.673	0.047		Ensemble	pr_wt_lr_all	0.607	0.038
patient-attentiveness	Raw	llama_fs	0.577	0.033	patient-warmth	Raw	llama_fscot	0.582	0.042
	LOGO	LOGO	0.589	0.030		LOGO	LOGO	0.564	0.060
	Ensemble	pr_mac_lr_top3	0.616	0.032		Ensemble	pr_wt_dtc_top2	0.582	0.030
patient-engagement	Raw	llama_fs	0.599	0.047	patient-empathy	Raw	llama_cot	0.589	0.067
	LOGO	LOGO	0.554	0.040		LOGO	LOGO	0.464	0.180
	Ensemble	pr_wt_dtc_top5	0.627	0.024		Ensemble	pr_mac_dtc_top5	0.616	0.086
patient-respect	Raw	llama_fs	0.546	0.037	patient-interactivity	Raw	flan_fs	0.544	0.035
	LOGO	LOGO	0.575	0.100		LOGO	LOGO	0.599	0.130
	Ensemble	pr_wt_dtc_top3	0.754	0.083		Ensemble	pr_mac_dtc_all	0.626	0.047

Table G1: Best-performing configurations for each social signal. For each signal, we report the highest balanced accuracy across three approaches: Raw (best individual model–prompt configuration), LOGO (leave-one-group-out logistic regression without weighting), and Ensemble (agreement-weighted ensemble with best aggregator and model selection). Method abbreviations: *flan/gemma/llama* = model type; *zs/fs/cot/fscot* = prompting strategy (zero-shot, few-shot, chain-of-thought, few-shot with CoT); *pr_wt/pr_mac* = weighted precision or macro-averaged precision confidence weighting; *lr/dtc* = logistic regression or decision tree classifier aggregator; *all/g/l3p1/top2/top3/top5* = model selection strategy (all models, Gemma only, LLaMA with all prompts, or top-*k* models by agreement). Bold indicates the best score for each signal.