LLM-based Agent Simulation for Maternal Health Interventions: Uncertainty Estimation and Decision-focused Evaluation

Sarah Martinson
Harvard University
sarahmartinson@g.harvard.edu

Lingkai Kong
Harvard University
lingkaikong@g.harvard.edu

Cheol Woo Kim
Harvard University
cwkim@g.harvard.edu

Aparna Taneja
Google Research India
aparnataneja@google.com

Milind Tambe
Harvard University
milind_tambe@harvard.edu

ABSTRACT

Agent-based simulation is crucial for modeling complex human behavior, yet traditional approaches require extensive domain knowledge and large datasets. In data-scarce healthcare settings where historic and counterfactual data are limited, large language models (LLMs) offer a promising alternative by leveraging broad world knowledge. This study examines an LLM-driven simulation of a maternal mobile health program, predicting beneficiaries' listening behavior when they receive health information via automated messages (control) or live representatives (intervention). Since uncertainty quantification is critical for decision-making in health interventions, we propose an LLM epistemic uncertainty estimation method based on binary entropy across multiple samples. We enhance model robustness through ensemble approaches, improving F1 score and model calibration compared to individual models. Beyond direct evaluation, we take a decision-focused approach, demonstrating how LLM predictions inform intervention feasibility and trial implementation in data-limited settings. The proposed method extends to public health, disaster response, and other domains requiring rapid intervention assessment under severe data constraints. All code and prompts used for this work can be found at https://github.com/sarahmart/LLM-ABS-ARMMAN-prediction.

KEYWORDS

LLM prediction, agent-based modeling, epistemic uncertainty-based aggregation, maternal health

ACM Reference Format:
Sarah Martinson, Lingkai Kong, Cheol Woo Kim, Aparna Taneja, and Milind Tambe. 2025. LLM-based Agent Simulation for Maternal Health Interventions: Uncertainty Estimation and Decision-focused Evaluation. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

Developing and deploying effective healthcare interventions in under-served regions often requires substantial time and resource investments. However, the absence of historical data or prior evaluations often makes it difficult to assess the efficacy of new health programs. Health workers and program managers may identify promising interventions, but the high costs of running trials or large-scale implementations complicate prioritization. To address this, we explore the potential of large language models (LLMs) for early-stage intervention assessment in data-deficient contexts, where behavioral data is unavailable and decision-making relies solely on contextual, population, and demographic information.

LLMs encode vast amounts of world knowledge, allowing them to approximate counterfactual agent-based predictions of individual behavior under hypothetical interventions. While these predictions do not constitute causal counterfactuals, they can provide plausible behavioral simulations based on encoded sociodemographic priors [34]. Our framework leverages these capabilities to assess maternal health programs before implementation.

Our study focuses on ARMMAN, a maternal health non-governmental organization in India that operates the mMitra program [4], an automated weekly call service delivering health messages to pregnant women and new mothers. While the programme has demonstrated success in improving health outcomes, sustaining engagement remains challenging [4]. Live service calls by health workers are used to support high-risk mothers, but resource constraints necessitate precise targeting. Currently, this targeting process follows a two-stage predictive framework: first, training a model to predict listening behavior, then using a Restless Multi-Armed Bandit [31] to allocate resources. While this approach effectively assigns live service calls, it requires extensive historical training data. There is no solution for low-data settings, such as evaluating interventions in new locations or under novel conditions. To explore whether LLM-powered agent-based predictions can provide robust appraisals in such settings, this study reimagines the existing mMitra program as an intervention *not yet implemented*. Given contextual details about the program and sociodemographic characteristics of potential participants, we prompt an LLM acting as a mother (agent) to predict binary engagement.

Different LLMs possess varying world knowledge, make predictions with different levels of uncertainty, and encode distinct biases. To enhance robustness and model calibration, we compare multiple LLMs and ensemble their predictions. Inspired by team-based agent collaboration frameworks [18, 19], we evaluate three ensembling strategies: direct averaging, epistemic uncertainty-weighted aggregation, and lowest-uncertainty prediction selection. We investigate their ability to balance robustness and calibration while mitigating biases in individual models, and assess their effectiveness in terms of predictive accuracy, F1 score, and calibration. Additionally,

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). This work is licenced under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence.

we adopt a decision-focused approach to evaluate how well LLM-based predictions support real-world decisions about whether to implement interventions in data-deficient contexts.

Contributions

We (1) introduce an LLM-based approach for prioritizing interventions and estimating their impact in the absence of historical behavioral data, and when real-world experimentation is limited or infeasible; (2) comprehensively compare and evaluate multiple ensembling methods for LLM prediction, demonstrating their effects on accuracy, F1 score, and log-likelihood; and (3) introduce a decision-focused evaluation pipeline leveraging LLM predictions and counterfactual modeling to guide intervention assessment in resource-constrained social good settings. While our primary focus is maternal health, the framework is generalizable to other domains requiring rapid, data-efficient decision-making, such as disaster response, pandemic control, and targeted social programs.

2 RELATED WORK

Agent-Based Modeling for Social Good. Agent-based modeling (ABM) has been widely used for public health, social welfare, and resource allocation [9, 35]. Successful applications of automated agents include pilot simulations for battlefield environments [27], airport evacuation modeling [29], training systems for disaster incident commanders [24], and task allocation and discovery for dynamic disaster response [20]. ABM has also demonstrated success in real implementations of maternal mobile health (mHealth) programs [30]. These models allow researchers and policymakers to experiment with virtual populations, identifying how different interventions might affect outcomes [9]. However, traditional ABMs typically rely on hand-crafted rules and historical time-series data to model agent behavior. In data-deficient settings, this dependence on dense behavioral records poses a significant challenge, as real-world datasets are often sparse or incomplete.

Recent advances in foundation models and LLMs suggest a new paradigm of simulation agents that encode extensive world knowledge [35]. Unlike handcrafted ABMs, LLM-based agents can generate counterfactual behavioral predictions even in low-data environments, offering more flexible and adaptive simulations. This capability is pertinent in AI for social impact contexts, where traditional modeling pipelines are often labor- and resource-intensive and domain-specific, limiting applicability and scalability [35]. Additionally, [5] discusses the integration of LLMs into ABM across various fields, including public health and social welfare. By incorporating LLMs, ABMs can simulate complex human behavior and interactions more effectively, even in data-sparse environments. However, limitations of LLM-based simulations, including ensuring their reliability and interpretability, and their real-world effectiveness remain underexplored [5]. Future research must validate these models through empirical studies to assess their reliability and robustness in real decision-making scenarios.

Predicting engagement in maternal healthcare initiatives relevant for large-scale mHealth programs [30]. Although models like Markovian restless bandits [30, 31] can optimize intervention resources, these approaches face limitations in non-Markovian contexts, particularly with multiple interventions or varied user behavior [30]. Moreover, standard ABMs offer little insight into counterfactual outcomes, as they rely on observed behaviour histories. In contrast, LLM agent simulations can hypothesize novel interactions by encoding prior world knowledge.

LLMs for Human Behavior Prediction. LLMs have demonstrated considerable potential for approximating human behavior by capturing language comprehension, cognitive heuristics, and common human systematic biases [8]. Recent studies have explored LLMs as proxies in social experiments, using value injection fine-tuning techniques to predict opinions [12], constructing simulacra that can produce personified responses for given characters [32], and demonstrating that LLMs perform better than traditional cognitive models in predicting human behavior in sequential decision-making tasks [21]. These advances hint at the potential for using LLMs to predict behavior of participants in health-related trials. However, challenges persist in calibrating LLM outputs, addressing overconfidence, and managing bias and variance [5].

LLM prediction ensembling. Previous work highlights advantages of including diverse agents in multi-agent teams, demonstrating that a collection of individually weaker agents can outperform uniform teams of individually superior agents [10, 18, 19]. Recent studies extend this to LLM ensembles, showing how techniques such as maximizing diversity [28], using sampling-and-voting methods at scale [16], and pairwise ranking with generative fusion [11] can yield notable performance gains over component LLM models. Building on these insights, we harness ensembling to stabilize and refine LLM-based behavior predictions, targeting applications in maternal health programs where reliable and diverse agent perspectives are important in guiding effective interventions.

Uncertainty Estimation of LLM predictions. Uncertainty quantification is vital in high-stakes settings where inaccurate predictions may lead to ineffective or even harmful outcomes. While accurate confidence estimates enable more reliable decision-making—where a model's certainty should be correlated with its correctness [33]—current LLM prediction methods often overlook uncertainty or fail to incorporate it systematically in ensembling [13]. Recent efforts distinguish epistemic uncertainty—reflecting gaps in model knowledge—from aleatoric uncertainty, arising from entropy in the underlying data distribution [1, 2]. These studies propose iterative prompting techniques to approximate uncertainties from LLM predictions. Since LLMs predict tokens from a vast textual corpus, they must manage both inherent randomness in language and data (aleatoric uncertainty), and gaps in their own knowledge (epistemic uncertainty). This suggests LLMs naturally contain internal representations of uncertainty that can be estimated to provide indications of model confidence [2].

3 METHOD

3.1 Data, Models & Prompting

Data. We have access to anonymized data from ARMMAN's mMitra program on 3000 mothers over 40 weeks. We consider two groups—a control group where all mothers receive only automated calls with health messages each week of the program, and an intervention group, where a random subset of mothers receives a live

call from a health worker instead of the automated message for a specific week. Intervention calls convey the same information as automated messages, tailored to the mother's stage of pregnancy or postpartum period. No mother receives an intervention more than once, and all intervention calls occur in the first six weeks.

For each mother, we consider binary actions—whether a mother received a live call or not—and corresponding continuous states—weekly listening times to health messages. To define engagement, listening times are converted to binary engagement states, where engagement corresponds to listening to a message for more than 30 seconds, while listening less (including non-answering) is considered unengaged for that week[1]. The LLMs' prediction task is to classify whether a mother will engage or not in a given week.

Each mother is associated with a set of numerically encoded sociodemographic features (Table 1). We assume contextual information about the program and a general population of mothers who may enroll is available. Specifically, we assume mothers are identified through maternal health clinic visits and provide participation consent. In our framework, LLMs act as mothers, making engagement predictions based on given sociodemographic profiles and program details, leveraging prior world knowledge about health interventions, telehealth adoption, and behavioral responses to generate predictions regarding a mother's engagement.

Models and hyper-parameters. For this study, we evaluate a selection of heavyweight and lightweight LLMs from Google, OpenAI, and Anthropic to capture a representative range of closed-source LLM capabilities. Heavyweight models include Gemini 1.5 Pro [7, 26] and GPT-4o [23], while lightweight models include Gemini 1.5 Flash [6], GPT-4o mini [22], and Claude Instant 1.2 [3]. Hyperparameters for each model are detailed in Table 2.

Prompting. We evaluate multiple intervention and control scenarios, described in Section 3.2. In each scenario, all models receive the same set of five prompts per mother per time step. Prompts vary slightly in wording but convey the same core information, including a description of the mHealth program, the sociodemographic characteristics of the mother, and a request for a prediction regarding her engagement at that time step.

All prompts state that simulation is weekly and specify the mode of message delivery—as a brief automated message or live service call from a health worker containing the same information. A sample prompt is provided in Box 1[2]. Characteristics are listed in Table 1, and key differences between control (orange) and intervention (green) prompts are highlighted. Since minor variations in prompt wording can affect LLM outputs [8, 25], we keep differences between intervention and control versions of each prompt minimal to maintain comparability. Each model is queried five times per prompt, yielding 25 predictions per mother per time step.

3.2 Simulation Scenarios

We examine three simulation settings: (1) *intervention*, (2) *counterfactual*, and (3) *control*. These allow us to evaluate intervention impact, compare predicted outcomes under counterfactual conditions, and establish a baseline for engagement without intervention.

[1]This aligns with engagement criteria used in previous evaluations of mMitra [30, 31].
[2]For the full set of prompts used in both intervention and control scenarios, see here.

Box 1: Sample LLM Prompt
<no intervention version><intervention version>

You are a mother enrolled in the ARMMAN Maternal and Child Healthcare Mobile Health program. ARMMAN is a non-governmental organization in India dedicated to reducing maternal and neonatal mortality among underprivileged communities. Through this program, you receive weekly preventive health information via brief <automated voice messages> <phone calls>.
In this simulation, each time step represents one week.
Below is your background and history with the program.
Your Background:

- You enrolled in the program during the `{enroll_gest_age}` week of your pregnancy.
- You are `{age_category}` years old.
- Your family's monthly income is INR `{income_bracket}`.
- Your education level is `{education_level}`.
- You speak `{language}`.
- You own a `{phone_owner}` phone.
- You prefer receiving calls during `{call_slot_preference}`.
- You enrolled in the program through `{channel_type}`.
- You are currently in the `{enroll_delivery_status}` stage.
- You have been pregnant `{g}` times, with `{p}` successful births.
- You have had `{s}` stillbirth(s) and have `{l}` living child(ren).

Past Behavior: The following is a record of your previous listening behavior (each representing one week): `{past_behavior}`
Based on this information, as well as the context of the program and on typical behavior of mothers in India, decide whether you will be engaged with the next automated health message.
Key Consideration: Engagement at a previous week does not necessarily imply engagement at the next, and lack of engagement at a previous week does not necessarily imply future lack of engagement. Engagement should depend on your specific circumstances that week (e.g. need for reassurance or information, phone availability, schedule, etc.). Being unable to answer a call that week implies a lack of engagement for that week.
Question: Will you be engaged with the next <automated health message><call from a health worker>?
Please respond with your final decision in the format: '##Yes##' for engagement or '##No##' for lack of engagement in this week. Your response need only contain one of the following: '##Yes##' OR '##No##'. No other text should be included.

Intervention. We select a sample with ∼ 60% of mothers receiving an intervention at some point during the 40-week period, reflecting real program constraints. This corresponds to ∼ 10% of the total population receiving a live call in each of the first six program weeks. This weekly threshold reflects a feasible allocation in resource-constrained programs, with an average of 0.015 live calls per mother over the entire population and program duration. LLMs simulating mothers who receive the intervention are prompted with the **intervention** prompt (Box 1) for the corresponding week. LLMs simulating mothers who do not receive the intervention are provided with the **no intervention** prompt, which specifies an automated telehealth message instead of a live call. At each time step, the LLM generates a binary engagement prediction.

We analyze two subsamples from this group: (1) a larger sample of 500 mothers, used to evaluate predictive performance of LLMs in an intervention context, and (2) a smaller representative subsample of 100 mothers, selected using K-means clustering to match the

Category	Available Characteristics
Enrollment Information	gestational age at enrollment, delivery status
Reproductive History	gravidity, parity, number of stillbirths, number of living children
Age Groups	<20, 20–24, 25–29, 30–34, 35+
Language	Hindi, Marathi, Kannada, Gujarati, English
Education Level	Illiterate, 1–5 years, 6–9 years, 10th pass, 12th pass, graduate, postgraduate
Phone Ownership	mother's phone, husband's phone, family phone
Preferred Call Time	8:30–10:30, 10:30–12:30, 12:30–15:30, 15:30–17:30, 17:30–19:30, 19:30–21:30
Enrollment Channel	community enrollment, hospital enrollment, ARMMAN enrollment
Monthly Income (INR)	0–5000, 5001–10000, 10001–15000, 15001–20000, 20001–25000, 25001–30000

Table 1: Sociodemographic characteristics available for mothers in the program.

Providor	Weight	Model	Generation Setup
Google	Heavy	Gemini 1.5 Pro	`model = gemini-1.5-pro-002, temperature = 1, max_tokens = 8192`
Google	Light	Gemini 1.5 Flash	`model = gemini-1.5-flash-002, temperature = 1, max_tokens = 8192`
OpenAI	Heavy	GPT-4o	`model = gpt-4o, temperature = 0.7, max_tokens = 2048`
OpenAI	Light	GPT-4o mini	`model = gpt-4o-mini, temperature = 0.7, max_tokens = 2048`
Anthropic	Light	Claude Instant 1.2	`model = claude-instant-v1, temperature = 0.7, max_tokens = 2048`

Table 2: Generating hyperparameters for various LLMs.

larger sample in terms of normalized feature values, mean engagement, and linear engagement trends. The smaller subsample is used to compare intervention effects across the three settings.

Counterfactual. We use the same representative subsample of 100 mothers as in the *intervention* group. In this scenario, all mothers receive **no intervention** prompts containing engagement history but omitting information regarding past interventions, allowing us to assess predicted engagement in the absence of explicit actions.

Control. This group consists of 100 distinct mothers who never received an intervention. All LLMs simulating these mothers receive **no intervention** prompts. However, in this case, action history is included and consists entirely of zero-action trajectories, reinforcing the absence of live calls.

3.3 Prediction Ensembling

For each simulation setting, we run all models in Table 2, generating $N = 25$ predictions per mother per time step. We compute mean predictions and associated epistemic uncertainty at each time step.

Epistemic Uncertainty. Epistemic uncertainty [15, 17] captures uncertainty in the model's knowledge and is distinct from aleatoric uncertainty, which arises from inherent randomness in data. Properly distinguishing between these sources of uncertainty allows us to weight predictions based on model confidence. Using N repeated binary predictions and methods from [14], we quantify predictive uncertainty for a given mother and time step as the binary entropy H of mean predictions across queries.

Let p_i be the individual predictions from a single model, and $\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i$ be their mean. Predictive uncertainty is then

$$H(\bar{p}) = -\bar{p}\log(\bar{p}) - (1-\bar{p})\log(1-\bar{p}).$$

Aleatoric uncertainty is the mean of individual prediction entropies

$$\frac{1}{N}\sum_{i=1}^{N} H(p_i) = \frac{1}{N}\sum_{i=1}^{N}[-p_i\log(p_i) - (1-p_i)\log(1-p_i)],$$

and epistemic uncertainty is then obtained as the difference

$$H(\bar{p}) - \frac{1}{N}\sum_{i=1}^{N} H(p_i).$$

This provides a per-mother, per-time-step uncertainty estimate.

Ensembling. We evaluate three ensembling strategies:

(1) Direct averaging. We compute mean predictions across models, presenting this average as the ensemble's final prediction probability (and binarizing where relevant). All models contribute equally to the ensemble, regardless of uncertainty.

(2) Uncertainty-weighted aggregation. We employ a Bayesian weighting mechanism using epistemic uncertainty to determine each model's contribution to the aggregated prediction. Lower epistemic uncertainties u_j for model j correspond to higher model precisions $\tau_j = \frac{1}{u_j}$ and greater influence in the aggregated result. For each time step, the combined prediction is a weighted average

$$p_{\text{combined}} = \frac{\sum_{j=1}^{M} \tau_j \bar{p}_j}{\sum_{j=1}^{M} \tau_j},$$

where $\bar{p}_j$ is the mean prediction from model j, τ_j is the corresponding precision, and M is the number of models in the ensemble. This ensures models with higher confidence exert greater influence on the final prediction.

Since different models may estimate epistemic uncertainty on different scales, we apply rank normalization across models before aggregation. For each model j with uncertainty estimate u_j , we compute the normalized uncertainty

$$u'_j = \frac{\text{Rank}(u_j)}{\max(\text{Rank}(u_1, u_2, \ldots, u_M))},$$

where $\text{Rank}(u_j)$ assigns a rank to each uncertainty relative to uncertainties of all M models. This preserves relative ordering while ensuring values are normalized to the same scale, preventing any single model's uncertainty from dominating the ensemble.

(3) Lowest-uncertainty selection. As a baseline, we use rank-normalized epistemic uncertainty values to select, for each mother at each time step, the model's prediction corresponding with the lowest uncertainty $p_{\text{selected}} = p_{\arg\min u_j}$.

3.4 Evaluation

Direct evaluation. For the larger group of 500 mothers, we simulate engagement trajectories over the full 40-week period for each LLM. This larger sample allows for a comprehensive assessment of model performance across diverse engagement trajectories. We compute three ensemble prediction values and conduct direct evaluations of component and ensemble predictions over time using accuracy, F1 score, and log-likelihood. Additionally, we perform a bias analysis for all component models and ensembles by breaking accuracy down by sociodemographic group (participant age, income, education level, and language).

Decision-focused Analysis. Once individual model and ensemble performance have been established, we prompt the LLMs for 15 weeks on the 100-mother subsamples of the *intervention*, *counterfactual*, and *control* groups. For each setting, we compute predicted engagement trajectories over time to determine whether the intervention setting exhibits increased engagement relative to no-intervention settings. Additionally, we analyze prediction transition probabilities over time to assess how LLMs model changes in retention and new/re-engagement following interventions.

4 EXPERIMENTAL RESULTS & ANALYSIS

Throughout this section, we compare performance of individual LLMs (Google's Gemini 1.5 Flash and 1.5 Pro, OpenAI's GPT-4o and GPT-4o mini, and Anthropic's Claude Instant) to the three ensemble methods described in Section 3.3: direct averaging (black), epistemic uncertainty-weighted aggregation (red), and lowest-uncertainty selection (gray). Ensembles are plotted for all five models, but curves are separated by provider for readability.

4.1 Evaluation Metrics

Here, we evaluate predictions of the LLMs for 500 mothers in the *intervention* setting over a 40-week period.

Accuracy. In Figure 1, we assess predictive accuracy of individual models and ensembling methods over time, with models grouped by provider. All models exhibit a decline in accuracy with time, which is expected in an autoregressive prediction setting because of error propagation from earlier predictions, making later weeks increasingly difficult to predict.

Claude Instant (bottom) consistently underperforms relative to other models, with accuracy exceeding 0.8 in only four weeks. By contrast, Gemini Flash and GPT-4o emerge as the strongest individual models in terms of accuracy. Gemini Flash consistently outperforms Gemini Pro, despite the latter being a more powerful model in general-purpose settings [7]. GPT-4o mini achieves strong early performance (~ 0.85 accuracy) but experiences greater fluctuation and a decline after ~ 30 weeks.

Notably, uncertainty-weighted aggregation (red) and direct averaging (black) mitigate this decline, indicating that ensemble methods help stabilize predictions when component model performance

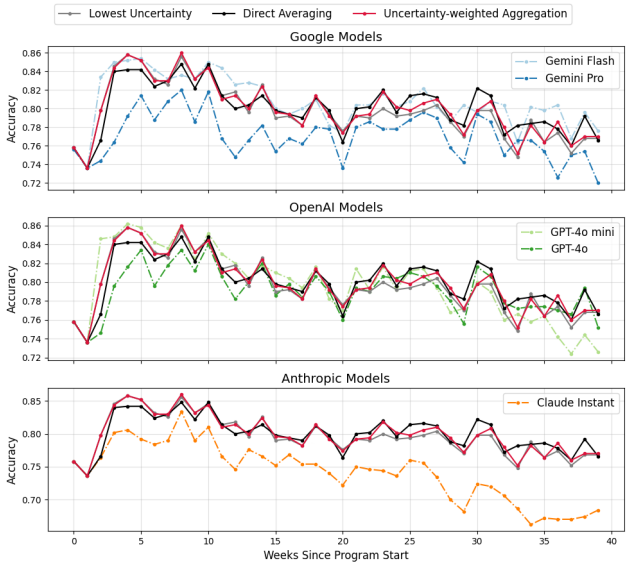

Figure 1: Mean accuracy of component and ensemble models over time, grouped by provider.

deteriorates. Meanwhile, GPT-4o exhibits a different trend, with initially lower accuracy that improves over time, surpassing GPT-4o mini after ~ 30 weeks. Ensemble methods effectively track these model performance shifts, dynamically adjusting to follow the best-performing model at each stage. Overall, ensemble methods provide robustness across models, improving predictions relative to weaker models such as Claude Instant and Gemini Pro, without sacrificing accuracy from stronger models.

4.1.1 F1 score. We plot F1 score (Figure 2) for all models over time for a more balanced evaluation to ensure models are not simply optimizing for the majority class. In this group, the total engagement proportion is 0.59, suggesting a possible class imbalance.

F1 scores demonstrate similar trends to accuracies. Unlike in accuracy results, Gemini Pro achieves slightly higher F1 scores than Gemini Flash. This suggests that Gemini Pro may have better recall, identifying more engagement cases, even at the cost of slightly reduced precision. GPT-4o and GPT-4o mini exhibit similar F1 score trends, while Claude Instant continues to demonstrate significantly lower performance compared to other models.

Most notably, no individual component model consistently outperforms either of the aggregation methods, reinforcing their effectiveness. By integrating predictions from different models, ensembling likely balances the precision-recall trade-off more effectively than any single model. Across all groups, uncertainty-weighted aggregation (red) and direct averaging (black) perform strongly, generally outperforming all individual models. In contrast, lowest-uncertainty selection (gray) underperforms slightly in the long term, reinforcing that selecting the single most confident prediction does not necessarily yield the best balance between precision and recall. The lower F1 scores for lowest-uncertainty selection suggest that high-confidence predictions might be biased toward precision, leading to reduced recall and classifier effectiveness.

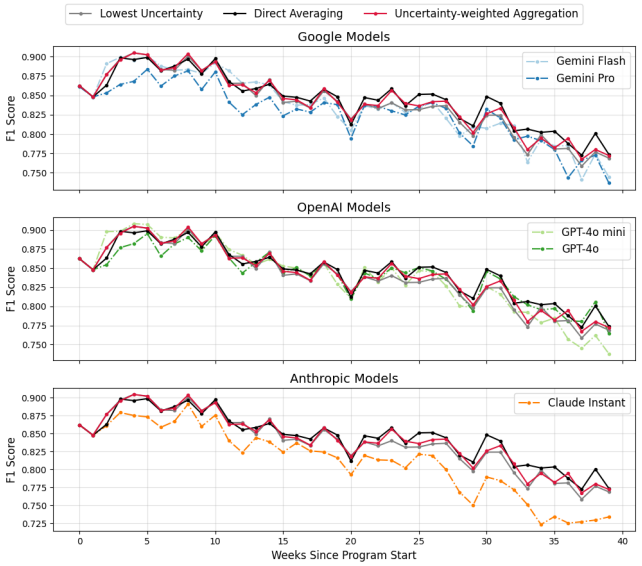

Figure 2: Mean F1 score of component and ensemble models over time, grouped by provider.

4.1.2 Log-likelihood. Figure 3 plots model log-likelihood over time, providing insight into model confidence and calibration. Unlike accuracy and F1 score, which assess binary correctness, log-likelihood captures both correctness and model confidence in predictions. Higher log-likelihoods indicate correct classifications *and* well-calibrated probability estimates, making it important for evaluating model reliability in uncertainty-aware settings such as ours.

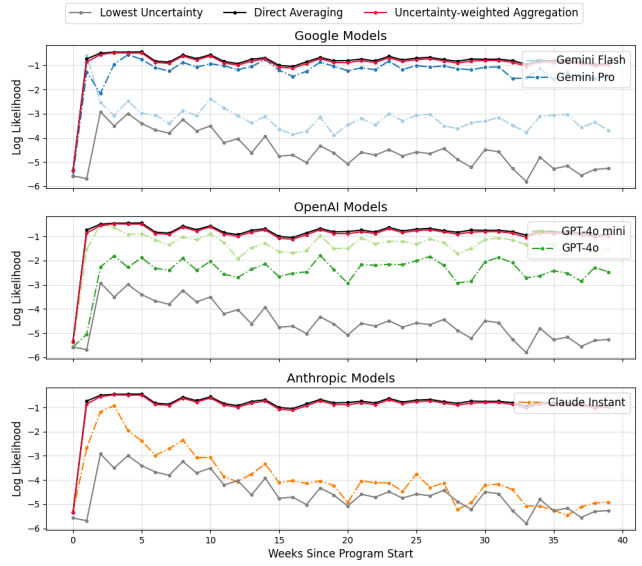

Figure 3: Mean log-likelihood of component and ensemble models over time, grouped by provider.

As with other metrics, log-likelihood exhibits an initial increase across all models during early calibration, followed by stabilization and a slight decline over time. Aggregation methods demonstrate highest log-likelihood values, indicating better alignment with true probabilities than component models and suggesting that aggregation mitigates overconfidence, leading to better-calibrated probability estimates and improved reliability.

Among individual models, Gemini Pro achieves significantly higher log-likelihood than Gemini Flash, despite having lower accuracy. A similar trend is observed with GPT-4o mini, which outperforms GPT-4o in log-likelihood despite lower raw accuracy. This suggests Gemini Pro and GPT-4o mini are better calibrated than their more accurate counterparts. While Gemini Flash and GPT-4o may achieve higher accuracy by making more confident predictions, their probability estimates may be less well-calibrated, leading to lower log-likelihood scores.

Among ensemble approaches, direct averaging (black) slightly outperforms uncertainty-weighted aggregation (red) in log-likelihood, both stabilizing around -0.75 in the long term. This suggests that weighting predictions by uncertainty may amplify miscalibrated models—an overconfident but systematically biased model may contribute higher precision weightings and lead to lower overall log-likelihood. Lowest-uncertainty selection (gray) performs particularly badly, indicating poor model calibration and an unreliable ensembling strategy—this method likely systematically selects overconfident predictions, even when incorrect.

4.1.3 Bias Analysis. We assess model fairness by evaluating accuracy across sociodemographic groups. Figure 4 plots accuracy for all component models and ensembles across income, age, education, and language groups. Bias is quantified as the maximum observed accuracy difference across feature categories for each model.

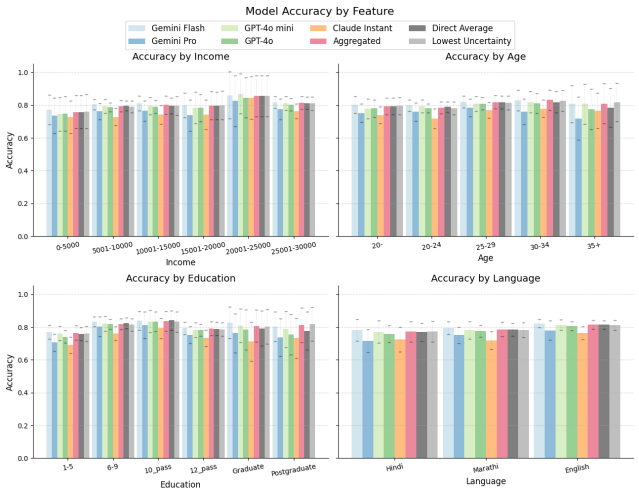

Figure 4: Average total accuracy by sociodemographic feature for individual and ensemble models.

Accuracy remains consistent across feature groups, demonstrating minimal model bias. There is a slight trend toward higher accuracy for higher income brackets, 10-pass and 12-pass education

categories, and English-speaking mothers, but no group exhibits significantly stronger performance than others, implying no strong systematic bias. Across sociodemographic groups, individual models exhibit similar accuracy distributions, with slight variations. Ensemble methods generally outperform individual models or at least match best performances, providing robust predictions across feature categories. The improved robustness of ensemble methods suggests aggregating predictions across models helps mitigate individual model biases and improve generalization.

4.2 Decision-focused Analysis

In this section, we focus on the two aggregation methods as they have demonstrated superior performance in accuracy, F1 score, and log-likelihood. Because of the poor performance of the lowest-uncertainty selection method, we exclude it from further analysis.

We analyze engagement predictions across the three settings using a cohort of 100 mothers over a 15-week period. We restrict analysis to this shorter time horizon, as engagement predictions become increasingly unstable beyond this point and accuracy tends to decline over time.

Total Engagement. Figure 5 presents the mean engagement proportion over time for each setting. Note the *counterfactual* setting (middle) does not include a ground truth curve, as no direct observations exist for this scenario.

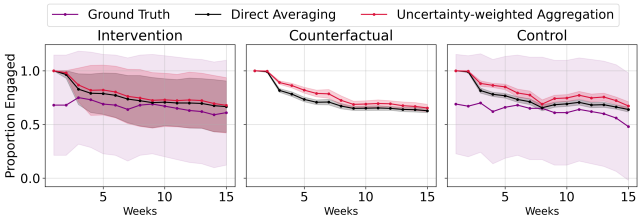

Figure 5: Mean engagement proportion over time in (left to right) the *intervention*, *counterfactual* and *control* settings.

Across settings, both aggregation methods (as well as individual component models, omitted for clarity) initially overpredict engagement with relatively low variance. Engagement declines over time in all cases, with uncertainty-weighted aggregation (red) consistently predicting slightly higher engagement than direct averaging (black). Both methods overestimate engagement relative to ground truth in *intervention* and *control* settings, particularly in the latter, suggesting systematic optimism in predictions. Despite initial bias, predictions gradually begin to align better with the observed engagement trend. In the *intervention* setting, this alignment occurs around the third week, while in the *control* setting, predictions this is closer to the sixth. Engagement predictions remain slightly elevated throughout the study period, indicating persistent overestimation of the ensemble methods.

Among the three settings, the *intervention* case exhibits the largest prediction variances, indicating greater uncertainty in engagement trajectories. The *counterfactual* engagement closely mirrors the *intervention* case. To quantify predicted effects of live call interventions, we compute differences in mean engagement between the *intervention* setting and the two no-intervention settings (Figure 6).

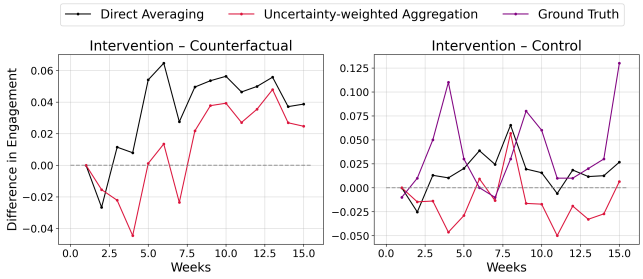

Figure 6: Difference in mean engagement over time between *intervention* and *counterfactual* settings (left), and *intervention* and *control* settings (right).

The difference between the *intervention* and *counterfactual* settings (left) is initially negative for both aggregation methods but increases over time. This suggests engagement in the intervention group does not immediately exceed counterfactual expectations, possibly due to model uncertainty in early predictions or a delay in the effect of live calls. However, in later weeks, engagement in the intervention setting surpasses the counterfactual trajectory, suggesting a growing intervention effect.

In contrast, the difference between *intervention* and *control* settings (right) follows more complex trajectories. Ground truth differences (purple) suggest an overall positive effect of live service call interventions, though engagement fluctuates over time. The high variability in the ground truth may reflect real-world fluctuations in engagement patterns, such as seasonal effects, external influences on participation, or heterogeneity in how different mothers respond to live calls. Model predictions do not fully capture this trend. While direct averaging (black) aligns more closely with the true improvement in engagement, uncertainty-weighted aggregation (red) tends to predict smaller differences. This likely arises because uncertainty-weighted aggregation overestimates engagement in the control setting (Figure 5), leading to an underestimation of relative intervention benefit. This may be due to the weighting mechanism amplifying predictions from models that are overconfident yet miscalibrated in control settings.

Transitions. For a more fine-grained analysis of the intervention's effects on engagement, we examine transition probabilities between 'engaged' and 'not engaged' states over time (Figure 7) to distinguish between retention (sustained engagement) and re-engagement (recovering previously disengaged users).

The *counterfactual* and *control* settings exhibit similar transition trends, suggesting that, in the absence of direct intervention, predicted engagement behavior follows a comparable trajectory across these settings, as expected. The similarity between *control* and *counterfactual* settings confirms that engagement trends without intervention are largely stable, validating that counterfactual predictions approximate a no-intervention scenario.

Across settings, aggregation method predictions tend to overestimate both engagement transitions ($0 \rightarrow 1$) and retention ($1 \rightarrow$

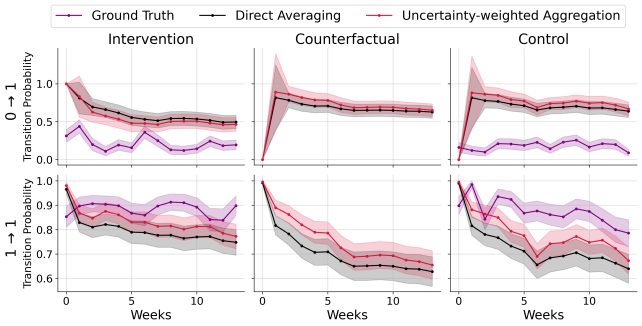

Figure 7: Transition probabilities over time for engagement states across *intervention*, *counterfactual*, and *control* settings. Top Row: Transitions from not engaged (0) to engaged (1). Bottom Row: Probability of remaining engaged (1 → 1).

1) relative to ground truth. However, in the *control* setting, models underestimate retention (1 → 1). In the *intervention* setting, the probability of transitioning from not engaged (0) to engaged (1) is lowest long-term among the three settings. While the intervention initially boosts engagement, its predicted effect on re-engagement diminishes over time. This suggests that live calls primarily help sustain engagement rather than recover disengaged users. Conversely, retention probabilities (1 → 1) are highest in the intervention setting, indicating that once engaged, users are more likely to stay engaged with live call interventions in early weeks. Considering both predictions and ground truth trends, service calls as an intervention appear to be more effective at sustaining engagement (1 → 1) than at driving new or re-engagement (0 → 1).

5 DISCUSSION AND CONCLUSIONS

Our findings suggest that LLMs can serve as effective predictive tools for engagement modeling in maternal health programs, particularly when combined through ensemble aggregation methods. However, our analysis also highlights key challenges, particularly regarding overconfidence in LLM predictions, which has been empirically observed in other settings [33].

While Gemini Flash performed well as an individual model in terms of accuracy and F1 score, aggregation methods demonstrated their strength in handling variability and uncertainty. These methods contribute to improved stability and generalization, especially in scenarios where individual models show greater fluctuations or overconfidence. F1 score of aggregated models is never outperformed by any component model, reinforcing benefits of ensemble methods in balancing precision and recall. Model aggregation has the greatest impact on log-likelihood, with direct averaging achieving slightly better log-likelihood values than uncertainty-weighted aggregation. This highlights the robustness of aggregation over any single model's prediction, improving probability calibration. Ensemble aggregation improves predictive robustness, particularly in settings with data sparsity, by leveraging model diversity.

Counterfactual predictions provide a valuable tool for intervention analysis, allowing us to simulate and compare engagement trends across settings. Findings indicate that interventions primarily sustain engagement (1 → 1) rather than drive re-engagement (0 → 1). Our results suggest LLMs can be a useful decision-support tool, but they require calibration and aggregation to mitigate overconfidence and ensure reliability. Specifically, direct application of individual LLM predictions may lead to biased intervention planning because of overconfident engagement forecasts.

Uncertainty-aware approaches should be further refined to improve decision-making in resource-limited settings, particularly by calibrating models to better distinguish between high-confidence and uncertain predictions.

5.1 Future Work

Main avenues for future work include: (1) Improving counterfactual modeling by investigating whether models with or without explicit knowledge of the intervention produce more reliable counterfactual predictions. (2) Exploring additional adaptive weighting schemes for ensemble methods, where model contributions change dynamically based on confidence *and* past performance. (3) Calibrating uncertainty estimates to ensure models more accurately reflect variability in engagement behavior. (4) Extending models to simulate program expansion over time, accounting for new participants entering the intervention, to provide a more realistic setting.

Our work demonstrates that LLMs, when properly aggregated, can provide meaningful engagement predictions to guide maternal health interventions. However, challenges related to uncertainty estimation, model calibration, and counterfactual prediction reliability remain key areas for future research. By addressing these limitations, LLM-driven approaches could play a significant role in scalable, data-efficient decision-making for social good programs.

Consent and Data Usage

Consent for participating in ARMMAN's mMitra program is received from all beneficiaries. All data collected through the program is owned by ARMMAN and only they are allowed to share data. This dataset will never be used by Google for any commercial purposes. All data used for this project was entirely anonymized before being parsed to any language model; no personally identifiable information is used. Data exchange and use was regulated through clearly defined exchange protocols including anonymization, read-access only to researchers, restricted use of the data for research purposes only, and approval by an ethics review committee registered with the Indian Council of Medical Research.

REFERENCES

[1] Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. 2024. To Believe or Not to Believe Your LLM: IterativePrompting for Estimating Epistemic Uncertainty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems.* https://openreview.net/forum?id=k6iyUfwdI9

[2] Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. 2024. Distinguishing the knowable from the unknowable with language models. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 22, 47 pages.

[3] Anthropic. 2024. Releasing Claude Instant 1.2. Retrieved Accessed: 2025-01-16 from https://www.anthropic.com/news/releasing-claude-instant-1-2

[4] ARMMAN. 2024. mMitra: A Free Mobile Voice Call Service for Pregnant Women and Mothers with Infants. https://armman.org/mmitra/

[5] Chao Gao, Xuwei Lan, Nan Li, et al. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications* 11 (2024), 1259. https://doi.org/10.1057/s41599-024-03611-3

[6] Google. 2024. Gemini 1.5 Flash Model Details. Retrieved Accessed: 2025-01-16 from https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-pro

[7] Google. 2024. Gemini 1.5 Pro Model Details. Retrieved Accessed: 2025-01-16 from https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-pro

[8] Rik Huijzer and Yannick Hill. 2023. Large Language Models Show Human Behavior. *OSF Preprints* (January 2023). https://doi.org/10.31234/osf.io/munc9

[9] Gauri Jain, Pradeep Varakantham, Haifeng Xu, Aparna Taneja, Prashant Doshi, and Milind Tambe. 2024. IRL for Restless Multi-armed Bandits with Applications in Maternal and Child Health. In *Pacific Rim International Conference on Artificial Intelligence*. Springer Nature Singapore, 165–178.

[10] Albert Jiang, Leandro Soriano Marcolino, Ariel D. Procaccia, Tuomas Sandholm, Nisarg Shah, and Milind Tambe. 2014. Diverse Randomized Agents Vote to Win. In *Advances in Neural Information Processing Systems 27 (NeurIPS 2014)*.

[11] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14165–14178. https://doi.org/10.18653/v1/2023.acl-long.792

[12] Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. 2023. From Values to Opinions: Predicting Human Behaviors and Stances Using Value-Injected Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 15539–15559. https://doi.org/10.18653/v1/2023.emnlp-main.961

[13] Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated Language Model Fine-Tuning for In-and Out-of-Distribution Data. In *Conference on Empirical Methods in Natural Language Processing*.

[14] Lingkai Kong, Harshavardhan Kamarthi, Peng Chen, B. Aditya Prakash, and Chao Zhang. 2023. KDD'23 Tutorial: Uncertainty Quantification in Deep Learning. Tutorial at KDD 2023. https://www.dropbox.com/scl/fo/gl1by69yjpjvpglabkh73/h?rlkey=m09c6h2yssrt2ubf7eq3j77bl&e=1&dl=0

[15] Lingkai Kong, Jimeng Sun, and Chao Zhang. 2020. SDE-Net: Equipping Deep Neural Networks with Uncertainty Estimates. In *International Conference on Machine Learning*. PMLR, 5405–5415.

[16] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More Agents Is All You Need. *Transactions on Machine Learning Research (TMLR)* (October 2024). https://openreview.net/forum?id=bgzUSZ8aeg Accepted by TMLR.

[17] Yinghao Li, Lingkai Kong, Yuanqi Du, Yue Yu, Yuchen Zhuang, Wenhao Mu, and Chao Zhang. [n.d.]. MUBen: Benchmarking the Uncertainty of Molecular Representation Models. *Transactions on Machine Learning Research* ([n. d.]).

[18] Leandro Soriano Marcolino, Albert Xin Jiang, and Milind Tambe. 2013. Multi-Agent Team Formation: Diversity Beats Strength?. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*. 279–285. https://dl.acm.org/doi/10.5555/2540128.2540170

[19] Leandro Soriano Marcolino, Haifeng Xu, Albert Xin Jiang, Milind Tambe, and Emma Bowring. 2014. Give a Hard Problem to a Diverse Team: Exploring Large Action Spaces. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. https://cdn.aaai.org/ojs/8880/8880-13-12408-1-2-20201228.pdf

[20] Ranjit Nair, Takayuki Ito, Milind Tambe, and Stacy Marsella. 2002. Task allocation in the robocup rescue simulation domain: A short note. In *RoboCup 2001: Robot Soccer World Cup V 5*. Springer, 751–754.

[21] Thuy Ngoc Nguyen, Kasturi Jamale, and Cleotilde Gonzalez. 2024. Predicting and Understanding Human Action Decisions: Insights from Large Language Models and Cognitive Instance-Based Learning. *arXiv preprint* (2024). https://arxiv.org/html/2407.09281v1

[22] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. Retrieved Accessed: 2025-01-16 from https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[23] OpenAI. 2024. Hello GPT-4o. Retrieved Accessed: 2025-01-16 from https://openai.com/index/hello-gpt-4o/

[24] Nathan Schurr, Pratik Patil, Fred Pighin, and Milind Tambe. 2006. Using multiagent teams to improve the training of incident commanders. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. 1490–1497.

[25] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=RIu5lyNXjT

[26] Sundar Pichai and Demis Hassabis. 2024. Our next-generation model: Gemini 1.5. Retrieved Accessed: 2025-01-16 from https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note

[27] Milind Tambe, W Lewis Johnson, Randolph M Jones, Frank Koss, John E Laird, Paul S Rosenbloom, and Karl Schwamb. 1995. Intelligent agents for interactive simulation environments. *AI magazine* 16, 1 (1995), 15–15.

[28] Selim Furkan Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. LLM-TOPLA: Efficient LLM Ensemble by Maximising Diversity. *arXiv preprint* arXiv:2410.03953 (October 2024). https://arxiv.org/abs/2410.03953

[29] Jason Tsai, Natalie Fridman, Emma Bowring, Matthew Brown, Shira Epstein, Gal A Kaminka, Stacy Marsella, Andrew Ogden, Inbal Rika, Ankur Sheel, et al. 2011. ESCAPES: evacuation simulation with children, authorities, parents, emotions, and social comparison.. In *AAMAS*, Vol. 11. Citeseer, 457–464.

[30] Shresth Verma, Arshika Lalan, Paula Rodriguez Diaz, Panayiotis Danassis, Amrita Mahale, Kumar Madhu Sudan, Aparna Hegde, Milind Tambe, and Aparna Taneja. 2024. Leveraging AI to Improve Health Information Access in the World's Largest Maternal Mobile Health Program. *AAAI* (December 2024). https://doi.org/10.1002/aaai.12206 Shresth Verma and Arshika Lalan are joint first authors.

[31] Shresth Verma, Aditya Mate, Kai Wang, Neha Madhiwalla, Aparna Hegde, Aparna Taneja, and Milind Tambe. 2023. Restless Multi-Armed Bandits for Maternal and Child Health: Results from Decision-Focused Learning. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. https://proceedings.neurips.cc/paper_files/paper/2014/file/8edd72158ccd2a879f79cb2538568fdc-Paper.pdf

[32] Qiujie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Yuejie Zhang, Rui Feng, and Shang Gao. 2024. Human Simulacra: A Step toward the Personification of Large Language Models. *arXiv preprint arXiv:2402.18180* (2024). arXiv:2402.18180 [cs.CY] https://arxiv.org/abs/2402.18180

[33] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=gjeQKFxFpZ

[34] Bo Yang, Jiaxian Guo, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Large Language Models as Theory of Mind Aware Generative Agents with Counterfactual Reflection. *arXiv preprint arXiv:2501.15355* (2024). arXiv:2501.15355 [cs.AI] https://arxiv.org/abs/2501.15355

[35] Yunfan Zhao, Niclas Boehmer, Aparna Taneja, and Milind Tambe. 2024. Towards Foundation-model-based Multiagent System to Accelerate AI for Social Impact. *arXiv preprint* arXiv:2412.07880 (December 2024). https://arxiv.org/abs/2412.07880