Towards Personalized Explanations for Health Simulations: A Mixed-Methods Framework for Stakeholder-Centric Summarization

Philippe J. Giabbanelli[1] **and Ameeta Agrawal**[2]
[1]Virginia Modeling, Analysis, and Simulation Center (VMASC), Old Dominion University
1030 University Blvd, Suffolk, VA 23435, USA
[2]Department of Computer Science, Portland State University
1900 SW 4th Ave, Portland, OR 97201, USA
pgiabban@odu.edu, ameeta@pdx.edu

Abstract

Modeling & Simulation (M&S) approaches such as agent-based models hold significant potential to support decision-making activities in health, with recent examples including the adoption of vaccines, and a vast literature on healthy eating behaviors and physical activity behaviors. These models are potentially usable by different stakeholder groups, as they support policy-makers to estimate the consequences of potential interventions and they can guide individuals in making healthy choices in complex environments. However, this potential may not be fully realized because of the models' complexity, which makes them inaccessible to the stakeholders who could benefit the most. While Large Language Models (LLMs) can translate simulation outputs and the design of models into text, current approaches typically rely on one-size-fits-all summaries that fail to reflect the varied informational needs and stylistic preferences of clinicians, policymakers, patients, caregivers, and health advocates. This limitation stems from a fundamental gap: we lack a systematic understanding of what these stakeholders need from explanations and how to tailor them accordingly. To address this gap, we present a step-by-step framework to identify stakeholder needs and guide LLMs in generating tailored explanations of health simulations. Our procedure uses a mixed-methods design by first eliciting the explanation needs and stylistic preferences of diverse health stakeholders, then optimizing the ability of LLMs to generate tailored outputs (e.g., via controllable attribute tuning), and then evaluating through a comprehensive range of metrics to further improve the tailored generation of summaries.

Introduction

Simulation models have been developed for many health application scenarios, such as finding the right treatment for depression given a *patient*'s characteristics (Wittenborn and Hosseinichimeh 2022), supporting *policymakers* in identifying efficient campaigns to promote healthy eating and reduce hypertension (Khademi et al. 2018), or optimizing the layout of a hospital (Dos Santos et al. 2025). These models can be complex: for example, the hypertension model accounts for the structure of social networks and the diffusion of social norms, as well as individual taste preferences and the eventual impact of food consumption patterns onto health outcomes. This complexity may create barriers to participation in the modeling process for non-modelers, particularly compounded with recurring issues related to model communication and low levels of traceability (Belfrage et al. 2024). These challenges may partly explain recent evidence from the literature on human-centered computing in which participants were primarily engaged in the early stage of model building but insufficiently as development progressed (Manellanga and David 2024). We consider that such lack of participation is a missed opportunity for validation (e.g., community members could compare the simulated journeys of agents to their own experiences) and can reduce buy-in. Ahrweiler et al. (2019) emphasized the importance of trust for buy-in: "the first and most important is that the clients want to understand the model[:] to trust results means to trust the process that produced them."

Given *(i)* the significant efforts devoted to building models and their demonstrated relevance for decision-making in the context of health and *(ii)* the challenges of engaging participants into the process to ensure the accuracy of findings and their translation into practices, many studies have proposed and evaluated means of engaging participants into the modeling process. Methods can vary based on criteria such as the targeted degree of participation, which ranges from providing information to co-deciding (Ferrand, Hassenforder, and Girard 2024). *We focus on the dissemination of results*, which is located as a low degree of participation.

As empirical studies on information dissemination and simulations have demonstrated, participants did not "want to look at a multitude of tables and scan through simulation results for interesting parameters; nor did they expect to watch the running model producing its results" (Ahrweiler et al. 2019) (Fig. 1). Researchers have co-developed interactive visualization environments with practitioners to explore models (Kammler et al. 2023), as exemplified by our work on policy-making and obesity in which policymakers could find key constructs and interrelationships in the model (Giabbanelli and Baniukiewicz 2018). However, ensuing usability studies showed that such platforms had a problematic learning curve or took too long to perform simple tasks (Giabbanelli and Vesuvala 2023). As a result, computational solutions to promote engagement with simulation models (particularly to explain a system) belong to two broad categories.

First, there are platforms that are intended to *facili-*

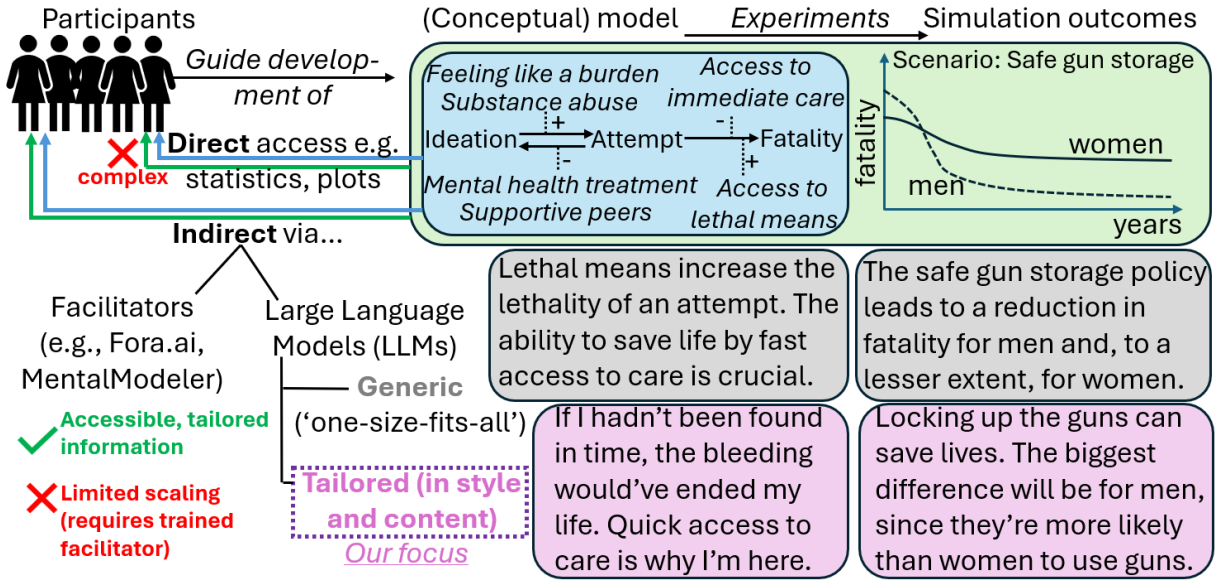

Figure 1: A model consists of elements and interrelationships from the problem domain, exemplified here as suicide prevention. The model can be programmed and then used for computational experiments, leading into simulation outcomes that can be differentiated by target populations. Our focus is on explaining models and simulations in a tailored manner (here for a lay audience) instead of the dominant and generic model-to-text translation via LLMs. This is more scalable than human facilitation.

tate engagement, typically within a workshop setting with a trained facilitator. `Fora.ai` allows participants to express the outcomes that they wish to see from the simulations (i.e., their 'concerns') and the results are primarily communicated on a map, given the platform's emphasis on environmental problems such as flooding (Zellner et al. 2025). `MentalModeler` supports participants in externalizing their mental models as Fuzzy Cognitive Map (visualized as node-and-link diagrams) and to see simulation outcomes in easily interpretable visuals showing which factors would increase or decrease (Gray et al. 2013). This tool has been used for over a decade, and the participatory technique of Fuzzy Cognitive Mapping has been abundantly used in health (Sarmiento et al. 2024) and medicine (Apostolopoulos et al. 2024). However, trained facilitators are not always available, so this approach is more appropriate for workshop settings where a small number of participants can be guided to learn new approaches than for large-scale deployment in which participants should independently access information. This scaling limitation echoes concerns on technology exclusion about who will be granted the benefits of new platforms for decision-making (Ahrweiler et al. 2025).

In contrast, the second approach uses technology to explain simulation results in familiar formats that people can interpret *independently*, which is the focus of this paper. This approach does not have facilitators and does not currently solicit feedback, unlike the platforms mentioned above. Rather, the focus is solely on information dissemination. Large Language Models (LLMs) are prominent in this space, as they can transform data into textual summaries (Fig. 1, grey boxes). LLMs and their uses have evolved rapidly to explain health simulations, starting the first work in 2022 that explained obesity and suicide models via GPT-3 (Shrestha et al. 2022). The focus was on decomposing large models so they could be 'fed' to GPT one piece at a time, leading to a focus on sentence-level generation. This early prototype suffered from both the fluency issues of GPT-3 (e.g., typos, grammatical mistakes) and the limitations of lossy decomposition algorithms, which removed some of the model's aspect. Progress on LLMs resulted in high fluency, coverage, and faithfulness scores, while advances in model-to-text generation produced paragraphs (rather than bags of sentences) with coherent themes and transitions (Gandee and Giabbanelli 2024). As technology matured, the scope was broadened from explaining the *structure* of a model (e.g., how do risk and preventive factors *generally* explain the transition from suicide ideation to attempt) to covering the *dynamics* of a simulation (e.g., how did a *simulated individual* receive mental health treatment after a non-lethal suicide attempt). For example, the latest framework focused on empathy to explain the simulated life of an individual (e.g., who they are, where they lived, how they dealt with stress), which provided a sense of immersion to readers based on a human study (Giabbanelli et al. 2025).

The emphasis has been on technical advancements in LLMs and in systems built around them, including prompts, retrieval augmented generation (RAG), and mixed-methods approaches. However, we argue that this progress has not yet resulted in achieving the fundamental mission of LLM-generated explanations for modeling and simulation in health: how do we communicate information that is understandable and actionable to each stakeholder? In particular, all studies on model-to-text translation via LLMs have taken a one-size-fits-all approach, but *communication should be*

tailored to address the varied needs of different stakeholders. Our vision is to move from one-size-fits-all AI-driven solutions to tailored communication that responds to individual needs, echoing some of the motivations found in the emerging field of participatory AI (Ahrweiler et al. 2025).

To appreciate why different people need different information, consider our introductory example in which a model serves to optimize the layout of a hospital. Hospital administrators need to estimate how the proposed layout would affect throughput and staffing needs, medical practitioners require operational logic (e.g., will the proximity of ICU to ER help to keep the average delay low?), patients and caregivers need to navigate the space, and public health officials need a layout that supports emergency preparedness requirements. In addition to different *content* needs, we also expect differences in *style*. For example, hospital administrators may need executive summaries with bullet points to support decision-making activities and a business-oriented language, whereas medical practitioners may favor terminology aligned with clinical workflows, and patients may appreciate an accessible and empathetic language.

The main contribution of this paper is a vision and framework for eliciting, incorporating, and evaluating the style and content needs of diverse stakeholders to obtain tailored information on health models and simulations.

The remainder of this paper is organized as follows. We provide a succinct background on the importance and the technical feasibility of generating tailored summaries of a model for different users and applications. Next, we propose a framework that identifies information needs (styles and content) from different stakeholder groups, then iteratively optimizes the alignment between LLM-generated summaries and the participants' needs. Given the growing importance of participatory AI, this paper ends with a brief discussion on how various forms of participation (from workshops to broader engagements) can contribute to design and evaluation of LLM-generated explanations.

Background

The Importance and Challenges of Transparency in Models and Simulations for Health

Models and simulations can support collaborative searches for solutions when complex problems are characterized by the needs for trade-offs, particularly monetary costs and differential impacts across populations. In addition, unintended consequences may be more readily identified and mitigated when groups of stakeholders examine the implications of a decision and how its modeling assumptions would perform in practice. For example, spatial analyses of obesity patterns show a close relationship between what people are exposed to (e.g., fast-food outlets near their home or workplace) and their weight (Patterson et al. 2025). Public health policies on childhood obesity, such as restricting the ability of new fast-food outlets to open near schools, are acceptable to young people (Savory et al. 2025) and face little opposition, especially where fast-foods are already common (Keeble et al. 2024). Models and simulations can help to design zoning policies, such as finding the minimal distance between fast-food outlets or with respect to schools, to achieve a target reduction in consumption over a simulated period (Baniukiewicz et al. 2018). However, planners note that businesses can change the business category to avoid regulation or open within other locations (Hassan et al. 2024), so unintended consequences include a *displacement* of outlets (e.g. to deprived areas) or a transformation of the food landscape that does not support healthy eating. If business owners, young people, planners, and obesity researchers could examine simulated zoning policies, they may note such consequences and either propose revisions to the model or identify pilot areas that are sufficiently constrained for the model to operate as intended.

Stakeholders' shared understanding of the problem and the novelty, concreteness, and richness of proposed solutions evolve alongside a model's degree of realism, but up to a certain point. While more realistic models can yield more detailed estimates or investigate complex tradeoffs, they can also overwhelm users, preventing them from acting on the insights they derived from the model, particularly within social contexts that exhibit strong power dynamics and favor prediction (Zellner et al. 2022). Simpler models may be better understood, resulting in faster analyses and improved implementations (Brooks and Tobias 1996), even if they do not provide the most accurate support for decision-making activities. There is thus evidence of the importance to balance representational fidelity and *end-user intelligibility*.

While there is abundant advice from researchers on using policy models for decision-making (Ghaffarzadegan, Lyneis, and Richardson 2011), actual evidence of policymakers using simulation models to inform policy choices is limited (Ahrweiler et al. 2019), e.g. to unique health challenges such as pandemic planning (Janssen and Helbig 2018; Haddad and Bugarin 2020). As the lack of end-user intelligibility is a key challenge, improving model transparency may enhance the likelihood of producing models that stakeholders can use.

A model may be better understood once simplified, either by removing inappropriate complexity (e.g., redundant variables) or by manually performing subtle alterations (e.g., exclude infrequent events, replacing feedback loops by constant) and checking that results between a simplified model align with its original version (Robinson and Brooks 2024). However, there are cautionary tales in the case of empirically-grounded individual level models such as Agent-Based Models (ABMs), which are of particular interest for health simulations. As noted by Sun et al. (2016), ABMs are generally created to provide predictions based on the specific needs of users at specific places, so oversimplification can decrease a model's usefulness, validity, and credibility (van der Zee 2017). For example, modelers have expressed concerns that simple models are usually achieved by ignoring spatial heterogeneity and individual variability (van der Zee 2017), which would overlook how policies work differently across populations (Yu 2024) and places (e.g., access to care, exposure to risk factors). In addition, when critical aspects are missed due to oversimplification, users may resort to their motivational biases to fill-in for missing information, for example by perceiving events

as more or less likely depending on whether they are desirable (Montibeller 2018). This can result in choosing a policy option based on biases rather than scientific models.

Generating Text to Address the Need for Simulation Transparency

Transparency does not only depend on the model: it also depends on the users. Some aspects of a model may be well-understood by one set of users because it relates to their lived experiences, while other aspects may quickly create a cognitive overload. As summarized by Robinson and Brooks (2024), "a user's comprehension of [a model is] subjective, dependent on the perceptions of the model users, and their knowledge and experience". Communicating about models thus requires an assessment of the target users, since their different mental models, expectations, and experiences bring 'subjective and active elements' into the modeling exercise (Hämäläinen et al. 2013).

When simulation results are communicated to modelers, summary statistics and data visualizations play an important role (St-Aubin et al. 2023). For instance, our simulation platform for suicide prevention (developed with the CDC) provides health outcomes (suicide ideation, attempt, fatality) change over time across gender or race and ethnicity (Huddleston et al. 2022). Users can examine *why* these outcomes happened by tracking the fraction of agents with specific factors such as being bullied, hopeless, or a victim of physical or sexual abuse. Results are provided as interactive visualizations as well as raw data that can be exported for statistical analyses or parsed by a screen reader for users with visual impairments. However, data visualizations and statistics are not suitable for every audience and/or they do not easily integrate in every workflow. There is thus significant research interest in complementing or replacing these artifacts by using LLMs to translate simulations into textual explanations (Fedeli and Manrique Negrin 2024; Dolha and Buchmann 2024; Fahland et al. 2024).

Recent works have considered that four parts of a simulation could be explained as text: functionalities (domain-specific knowledge about the purpose and capabilities of the simulation), validity (metrics to evaluate in which situations and to which extent the results can be trusted), architecture (what inputs parameters can be modified and how do they affect observable outputs), and operations (libraries and software components involved during execution) (Fedeli and Manrique Negrin 2024). So far, *no study has examined which aspects of a simulation model should be conveyed to a given audience, in which style, or how to measure the reactions of an audience to textual summaries of simulations*.

Tailoring Text Generation by LLMs for Different Users and Applications

Models can involve many parameters, rules, datasets for calibration and validation, and simulation scenarios. As explained above, detailing the minute aspects of every component to every user would not be productive. Rather, we need to *summarize* the aspects of a model that matter to an individual (content) in a manner that supports their decision-making activities (style). LLMs are now widely used to generate summaries (Olabisi and Agrawal 2024), with several reviews detailing applications to clinical settings (Bednarczyk et al. 2025; Shool et al. 2025). Concerns about using LLMs in healthcare often focus on hallucinations, biased or stereotyped outputs (Omar et al. 2025), and inconsistent reasoning across languages (Schlicht et al. 2025) or multiple uses of the same prompt. These risks are especially important to monitor when patient safety is at stake, and robust oversight during deployment is essential. However, demanding absolute perfection from LLMs may be counterproductive. Rejecting systems for occasional imperfections can limit access to information (e.g., end-users cannot interpret simulation results without assistance) or force reliance on human intermediaries, such as modelers, who are neither perfectly accurate nor always available. Notably, a recent study found that physicians preferred LLM-generated answers over those from other physicians on eight of nine clinical axes (Singhal et al. 2025). In another study, physicians found 81% of LLM-generated summaries equivalent or superior to summaries from medical experts on completeness, correctness, and conciseness (Van Veen et al. 2024).

Given that LLMs can produce good summaries in health, the next step is to tailor these summaries to the needs and intents of individual users. For example, a generic description of scientific simulations may cover both the model and the domain, thus non-modelers would struggle with domain-specific terminology while subject matter experts may struggle with simulation techniques – as a result neither audience is satisfied. Consequently, *customizable summarization* is essential to accommodate individual user preferences (Wang et al. 2025). Adapting to a user's needs relies on controllable attributes (Urlana et al. 2023), such as summary length (to produce executive reports), writing style (to match the desired tone or message), information coverage (to include essential content), content diversity (increases the variety of topics covered), and topic control (ensures clarity by focusing on specific themes in research papers or reports). For instance, style-controlled summaries can maintain a consistent tone across different communication channels, while topic-controlled summaries enhance coherence by emphasizing specific areas of interest. Despite the technical feasibility of creating tailored summaries, this has *never been explored in the context of facilitating access to scientific simulations since the needs of each user group have not been assessed*.

Proposed Framework

Our proposal has two broad steps, detailed in the following two subsections and summarized in Figures 2 and 3. In the first step, we create technically correct summaries of a model's structure and simulation outcomes (as assessed by modelers) and we use them to measure the reactions of participants across groups to summaries. Given these needs, the second step generates several summaries to cover different potential preferences in content and style, and we re-assess them with participants for iterative feedback.

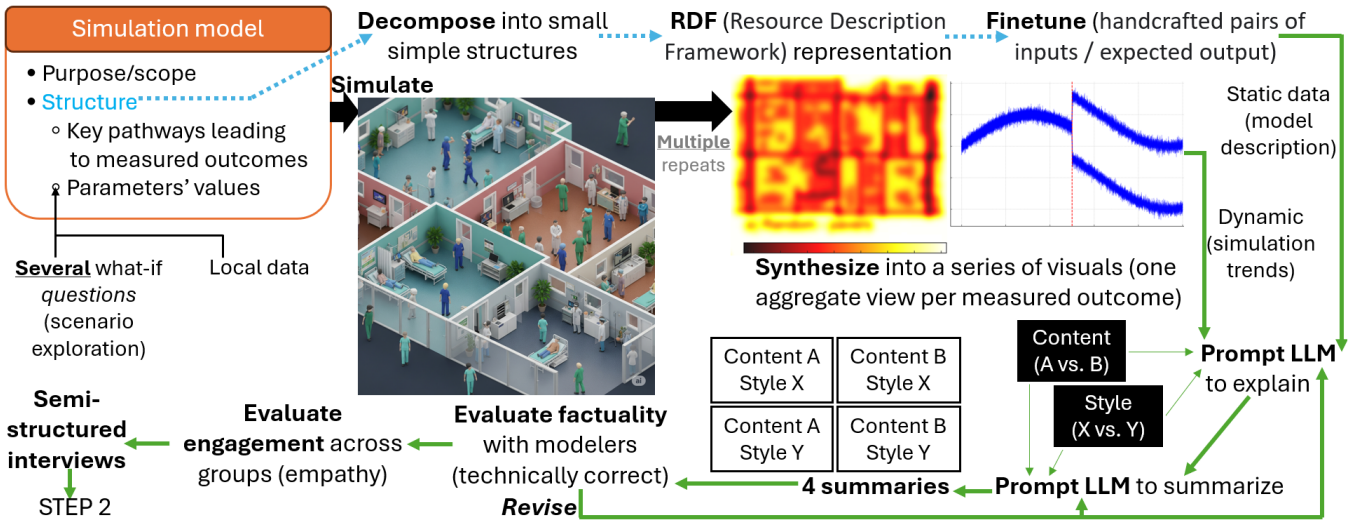

Figure 2: A simulation model has (1) a static structure, which can be decomposed and transformed into text using existing solutions (Gandee and Giabbanelli 2024), and (2) dynamic datasets (e.g., the states of simulated agents over time) created in response to simulation scenarios. Static and dynamic elements are turned into summaries (as a starting point) and evaluated.

Step 1: Identify Information Needs and Preferred Styles From Different Stakeholder Groups

A simulation model has two parts: a *static part that describes the model* (e.g., purpose, scope of what is included/excluded, main parameters and how they drive outputs) and *a dynamic dataset generated through simulations*. Translating the static part proceeds through three sub-steps.

① First, we split the full specification of the model into smaller parts that collectively contain the information (i.e., decomposition facilitates information processing but does not lose information). Available algorithms include 'RDF Walks' (Mussa et al. 2024) and our hierarchical approach (Gandee and Giabbanelli 2024). Breaking down complex models into sufficiently small sizes helps LLMs such as GPT in generating well-formed sentences (Shrestha et al. 2022), as measured by standard text quality scores in Natural Language Processing (e.g. ROUGE, BLEU). Note that the order produced by this *linearization* process (taking a graph and producing a sequential list) affects the LLM's performance, so the ordering can be optimized to preserve attribution and reduce hallucination. In addition, a hierarchical decomposition can help to structure the summary. Such decomposition shouldn't just be a technical inventory of model components, as it would be difficult to explain agents separately from physical spaces or processes. Rather, a well-formed decomposition should mirror a logical narrative that an explanation can follow, so the reader understands the model in a natural flow. For instance, a disease-progression perspective provides a useful organizing principle because it aligns with how both domain experts and lay readers think about the system: what happens first (who is in the population, what's their health status, how do they become exposed), what happens next (how an infection progresses over time), and how different processes connect (testing and diagnosis, isolation, treatment).

② Second, we represent each part in a structured manner. The smaller structures of a model consist of triplets, where one element has a relation with another element. The triplets can be represented in different ways, such as "head A — relation X — tail B" or via tags as <head>A</head><relate>X</relate><tail>B</tail>. Recent works showed that the choice of representation does have an impact with current LLMs, but it was not significant (He et al. 2025). We thus use RDF as an example in this section, since it has been studied in several other studies on LLMs for medical applications (Mavridis et al. 2025), but developers may use another representation.

③ Third, some health professionals use a web portal to access some LLM, type a specialized medical prompt, and find the answer unsatisfactory (Ponzo et al. 2024) – as would be expected in using a general purpose LLM. We emphasize that LLMs should not be expected to serve as oracles equipped with universal expertise: rather, the onus is on users to correctly employ LLMs as a component in an engineered pipeline. In particular, we must ensure domain adaptation through fine-tuning and/or retrieval augmented generation with authoritative medical sources. In our context, fine-tuning could be onerous by manually crafting a large number of pairs of <sample model input, expected textual output>. In an experimental study on two health cases (suicide and obesity), we explored the response curve between investing in more fine-tuning and the quality of the output (Giabbanelli et al. 2024a). We found that a handful of examples (few-shot learning) were sufficient for saturation, but they are also *necessary* as performances on zero-shot learning were inadequate.

The three steps above transform the structure of a model into text by feeding the entire structure to the LLM in small chunks ('bite size'). As we turn our attention to the translation of simulations to text, we face a different problem

of *volume*. A simulation may involve many virtual agents, whose activities span a long time, and many simulation runs may be needed to obtain a distribution of results (e.g., characterize health outcomes with a 95% confidence interval). Consequently, we cannot 'feed' the complete simulation results to the LLM in small increments in the same way as we decomposed the whole structure. A simple approach is to perform a statistical analysis and provide it as input to the LLM. In this case, the developer is responsible for deciding which trends are important and how to describe them in text for the LLM. Alternatively, we can visualize the simulation data and provide these visualizations as input to multi-modal LLMs (i.e., that can handle both text and images) such as OpenAI's GPT-4o or Google's Gemini. In this case, the developer uses several visualizations to reduce data while preserving spatial patterns (to know how agents interact with their environment) or variability (to express uncertainty). The visualizations can be advanced as they are intended for use by the LLM rather than by participants.

After combining the LLM-generated text generated for the complete model's structure and insights from the simulations, we can summarize based on controllable aspects: what a stakeholder wants to know (content or information coverage) and how it should be expressed (writing style). At this stage, we do not yet know their preferences so we only generate different *candidate* summaries to elicit feedback that will guide the next text generation. We thus recommend using designed experiments to generate several summaries based on content, style, and their interaction. For instance, if each controllable aspect is simplified by two options, then we have 2^2 factorial design, resulting in four summaries.

Before evaluating the summaries across stakeholders, we need to ensure correctness. There is little value in producing summaries that speak about the right topics in the desired style but it may support the wrong decisions based on a misleading interpretation of the model. Since evaluating summaries for factuality with crowdsourced workers may not be reliable, we recommend evaluating the summaries with modelers based on a questionnaire that covers three dimensions; whether each text contains factual errors (i.e., factuality labeling) and if so, explain the issues (error reasons) and categorize them (error types) as a knowledge error (hallucinated or inaccurate information), a reasoning error (flawed logic or reasoning), or irrelevant (content unrelated to the case). Each text should be evaluated by at least two modelers and their agreement rate can be calculated using weighted kappa to ensure reliability. If the score is unsatisfactory then the error must be located (was it the summarization? the model-to-text? the simulation-to-text?), the generation process revised, and the summaries re-evaluated.

Once the 'candidate' summaries are technically correct, we can pilot them with stakeholders to elicit and measure reactions depending on the specifics of a project. In the context of health, *empathy* is especially important as policy decisions impact vulnerable groups. Empathetic stories trigger emotional engagement, which motivates action-oriented decisions, thus supporting the translation of simulation outcomes into practice. Although there are several validated empathy questionnaires, time-efficient options are particularly valuable to maximize participation and response quality. Since perceiving a narrative as immersive and compelling depends on the mental state of the reader, we recommend using the validated Toronto Empathy Questionnaire (TEQ; 16 items) to obtain multiple empathy measures (Spreng et al. 2009). It has been used both for human readers who evaluate LLM-produced narratives (Shen et al. 2024) and to test an LLM's direct ability at producing empathetic text (Welivita and Pu 2024). In short, each participant would complete TEQ then receive four summaries (corresponding to the combinations of two controllable attributes) and complete a validated questionnaire for each one, such as the State Empathy Scale (12 items). We pilot-tested this protocol on measuring empathy from LLM-generated summaries with seven participants and noted a median response time of 19.16 minutes. The reasons behind participants' preferences and attitudes in the surveys can be further studied by one-on-one interviews.

Step 2: Optimize the Alignment of Language Models and Stakeholder Communication Needs

As the set of summaries were generated based on a design of experiments, a *factorial analysis* can decompose the participants' reactions as a function of the controllable aspects (information content and writing style). This analysis should be performed for each group of stakeholders, as the goal is to generate insight into the preferred modalities of each group. Note that a 'group of stakeholder' is not necessarily defined by role (e.g., patients, caregivers, physicians, healthcare administrator), so a complementary study may be needed to identify meaningful clusters (Lavin et al. 2018). Additional analyses can include effect sizes to indicate the magnitude of the observed effects, a post-hoc power analysis to examine whether the sample size was adequate, Cronbach's alpha to evaluate the internal consistency of each instrument within the specific population of respondents (as the questionnaire were validated in a more general sample), and repeated measures ANOVA to analyze how reactions vary across summaries. If qualitative one-on-one interviews were performed, then discourse analysis can examine how different groups use language, as the interviews may reveal differing narratives, framings, or ideological stances.

After completing the data analysis to identify the preferred features, we steer the LLM to generate new summaries that match the preferred content and styles of each group. *User-centric summarization* is challenging as people have long struggled to determine what is a good summary. Useful summaries depend on three types of contextual factors: *input factors* (the material that will be summarized), *purpose factors* (the intended purpose of the summary), and *output factors* (the characteristics of the generated summary).

Recent advances in text summarization encompass a spectrum of approaches (extractive, abstractive, generative, emerging hybrid architectures), each offering unique strengths. Extractive summarization selects salient sentences or phrases directly from the source text, maximizing factual retention and minimizing distortion, whereas abstractive methods rephrase and restructure ideas to produce more

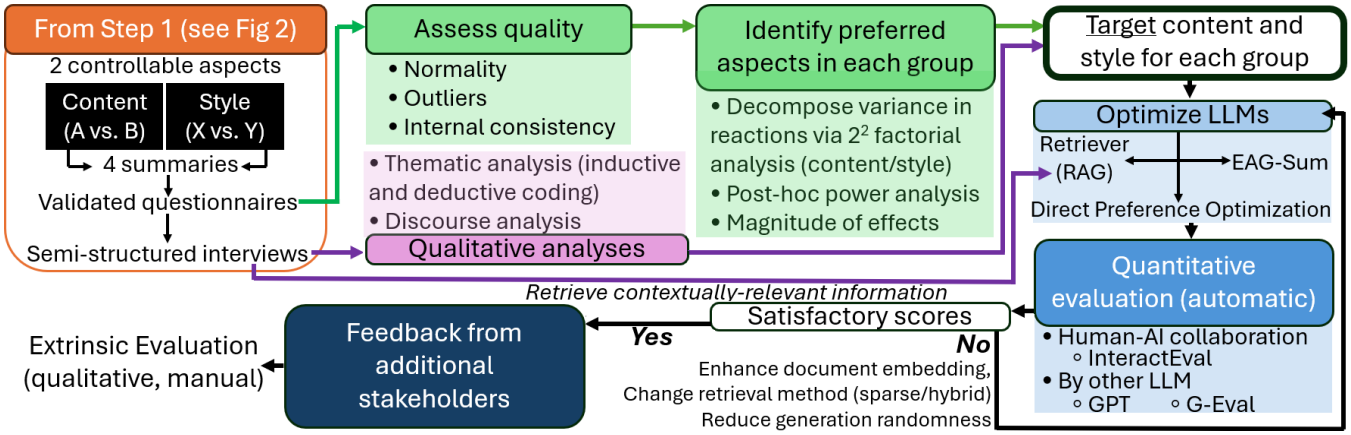

Figure 3: Our second step analyses the survey data to find what each group needs in a summary. Then, we steer LLMs in producing summaries that match these needs. An optimization process is involved, as the new summaries should be automatically assessed and the architecture adjusted if the scores are insufficient. Finally, the results can be presented to stakeholders.

natural, human-like summaries. Generative models, particularly those built on transformer architectures, have emerged as powerful tools capable of synthesizing information across multiple documents and producing contextually rich narratives, though sometimes at the expense of factual precision.

Ahmed and Hemanth (2025) demonstrate that extractive models excel in factual accuracy, abstractive models achieve greater coherence and conciseness, and generative models capture subtle contextual nuances. Their proposed `EAG-Sum` hybrid framework strategically integrates all three paradigms in a multi-stage process: (1) generating an extractive "skeleton" summary with a transformer-based model such as `BERTSum` to ensure coverage of core factual details; (2) refining this skeleton with an abstractive model like `PEGASUS` or `BART` to improve fluency, reduce redundancy, and introduce novel sentence structures; and (3) applying generative contextual enhancement via a model such as GPT-series to adapt tone, style, and domain specificity, adding auxiliary detail where needed.

There remains significant scope for improvement in steering LLMs toward accurate and desirable outputs through domain adaptation and human feedback-driven alignment strategies. Traditional supervised fine-tuning (SFT) alone is insufficient, as it optimizes against static gold-standard labels and fails to capture the richness of human preferences in open-ended generation. Recent approaches instead treat alignment as a preference optimization problem, where models are adapted to approximate the behavior preferred by human evaluators. Reinforcement Learning from Human Feedback (RLHF) has become a standard paradigm: a reward model is trained on human preference comparisons, and the base language model is then fine-tuned via reinforcement learning to maximize expected reward. Direct Preference Optimization (DPO) (Rafailov et al. 2023) reformulates alignment by removing the explicit reward model. Instead, it optimizes the language model parameters to directly maximize the log-likelihood ratio of preferred vs. dispreferred responses, using pairwise preference data. Diverse AI Feedback (Yu et al. 2025) goes beyond pairwise preferences by incorporating heterogeneous forms of supervisory signals into the optimization objective. Specifically, it integrates: (1) critique feedback, which provides structured error annotations and diagnostic signals; (2) refinement feedback, in which annotators (or auxiliary models) propose partial rewrites at the span or sentence level; and (3) ranking-based preferences, which offer global desirability signals across candidate completions. Using proven and integrated solutions can simplify the process instead of manually combining some of the (slightly) older tools.

While the approaches mentioned above provide a comprehensive set of tools to generate summaries, there is no guarantee that they will be optimal at first. For example, recent research has exposed counterintuitive pitfalls: Peters and Chin-Yee (2025) show that explicitly instructing LLMs to produce more faithful summaries can backfire, increasing overgeneralization rates by up to 15% in some models. Assessment is thus needed and may be followed by mitigation strategies such as changing some of the parameters of the RAG (Figure 3) or reducing generation randomness (e.g., lowering temperature to 0). The assessment should ensure that the newly generated summaries continue to logically follow from the content that was approved by modelers for factuality. While methodological advances have expanded the capabilities of summarization systems, their evaluation remains a critical and evolving challenge. Evaluating a large number of candidate summaries with respect to the preferences of each group would become time-consuming for human readers and it may not be mindful of the time commitment of participants. Thus, LLMs can be used to perform some of the evaluation. Studies have shown that GPT produced better preference and factuality ratings than conventional evaluation metrics on several datasets (Gao et al. 2023). Multi-dimensional evaluation metrics such as UniEval (Zhong et al. 2022) correlate more strongly with human judgments. More recently, innovative *human–AI collaborative evaluation* approaches, such as In-

teractEval (Chu et al. 2025), combine the high-level reasoning and flexibility of human Think-Aloud protocols with the consistency and breadth of LLM-generated checklists, producing superior benchmark performance on summarization datasets. The analysis by Chu et al. (2025) reveals that humans excel at identifying internal quality attributes (coherence, fluency), while LLMs better capture external alignment (consistency, relevance), suggesting the value of integrating both perspectives.

Once optimized summaries have been produced for each stakeholder group, the next step is to share them back with participants. This step must be approached with the understanding that stakeholders are not simply a source of data for iterative model refinement. They are often busy professionals or community members who volunteer their time because they care about solving a real problem. If we ask them to engage again to provide feedback on improved LLM-generated summaries, the interaction should deliver tangible value to them as well. We therefore recommend presenting the results in formats that benefit participants, such as workshops or educational sessions, where they can both learn from the findings and connect with other stakeholders.

Discussion: On Participatory AI

Given the importance of engaging different groups of users with modeling and simulation, we articulated a vision and complete process that leverages advances in (multimodal) LLMs to produce summaries tailored to the informational needs and styles of each group. Our framework is intentionally broad and mixed-methods, as we need several measures to capture the needs of participants and the extent to which a summary addresses these needs. The scientific basis for this framework will be strengthened by collecting experimental data, performing ablation studies to measure the effect of each part of the framework (which may result in a simplified framework), and comparing strategies (e.g., few-shot learning vs. supervised fine-tuning vs. preference optimization). While our vision focused on measures related to content and style, there are also *operational metrics*. If modelers visit low-resource settings where connectivity and hardware are constrained, they still need the LLM to generate text in a timely manner (e.g., measure response time). This may call for architectures centered on lightweight open-source LLMs and/or hybrid pipelines with edge computing to push queries on a user's hardware (to the extent possible) and only depend on the cloud for occasional secure retrieval.

In practice, stakeholders rarely read a summary, instantly trust its conclusions, and know exactly how to act. They may question the assumptions behind the model, the impact of alternative scenarios, or the reasons for specific simulation outputs. Provenance information is critical for answering these questions (Gierend et al. 2024), especially when results are counter-intuitive or central to justifying an intervention. In our context, however, provenance is complex: generated summaries can vary with LLM parameters (e.g., temperature, stochastic routing in mixture-of-experts models), the order of graph linearization, or the retrieval index used. This leads to a human-factors challenge: we are asking participants to trust explanations generated by an LLM (a technology they may already distrust) in order to build trust in a simulation model, which may also mistrusted. As Hinrichs et al. (2025) note in the context of mental health, "barriers to the integration of AI primarily stem from issues related to trust and confidence in the system, end-user acceptance, and system transparency". While this trust paradox is real, the alternative is less desirable, since continuing with the current status quo would mean leaving participants to face significant barriers while engaging with modeling.

Going forward, we thus envision the creation of more interactive environments that *extend* the textual summaries discussed in this paper. This immediately raises the question: *how* would individuals interact with the content? For example, the field of *visual analytics* has long been applied to healthcare and it supports interactions through details-on-demand on linked visualizations. In our prototype, a part of the summary can be expanded into complete paragraphs and the corresponding part of the model is provided as a node-and-link diagram (Gandee et al. 2024). However, the need to create text in this paper was motivated by the difficulty of engaging various audiences with scientific visualizations, thus embedding the text within overly technical platforms may defeat this purpose. An alternative could be to promote simpler *conversational* interactions by voice or text, but they raise numerous technical challenges to clarify a user's question and provide answers by combining information (Giabbanelli et al. 2024b).

Our framework articulated numerous qualitative and quantitative assessments to ensure that each group is presented with a summary that is factual and addresses their preferences. But we should not lose track of the bigger picture: the text is not the end goal for assessment. We create summaries to support decision-making activities in each group, so the ultimate demonstration that the pipeline works lies in its ability to affect decisions. We thus need to broaden evaluation to include downstream decision metrics: does a tailored summary change the decisions stakeholders make compared to a generic summary? A randomized controlled trial may assign stakeholders to different summaries (e.g., generic vs. tailored) to measure the impact of the decision.

References

Ahmed, R.; and Hemanth, D. J. 2025. Hybrid text summarization: Integrating extractive and abstractive models for enhanced cross-domain summarization. *Intelligent Decision Technologies*, 18724981251322745.

Ahrweiler, P.; et al. 2019. Co-designing social simulation models for policy advise: lessons learned from the INFSO-SKIN study. In *2019 Spring simulation conference (SpringSim)*, 1–12. IEEE.

Ahrweiler, P.; et al. 2025. Inclusive technology co-design for participatory AI. *Participatory Artificial Intelligence in Public Social Services: From Bias to Fairness in Assessing Beneficiaries*, 35–62.

Apostolopoulos, I. D.; et al. 2024. Fuzzy cognitive map applications in medicine over the last two decades: A review study. *Bioengineering*, 11(2): 139.

Baniukiewicz, M.; et al. 2018. Capturing the fast-food landscape in England using large-scale network analysis. *EPJ Data Science*, 7(1): 39.

Bednarczyk, L.; et al. 2025. Scientific evidence for clinical text summarization using large language models: scoping review. *J. Medical Internet Research*, 27: e68998.

Belfrage, M.; et al. 2024. Simulating change: A systematic literature review of agent-based models for policy-making. In *2024 Annual Modeling and Simulation Conference*, 1–13. IEEE.

Brooks, R. J.; and Tobias, A. M. 1996. Choosing the best model: Level of detail, complexity, and model performance. *Mathematical and computer modelling*, 24(4): 1–14.

Chu, S. Y.; et al. 2025. Think together and work better: Combining humans' and LLMs' think-aloud outcomes for effective text evaluation. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems*, 1–23.

Dolha, D. N.; and Buchmann, R. A. 2024. Generative AI for BPMN process analysis: experiments with multi-modal process representations. In *int. conf. on Business Informatics Research*, 19–35. Springer.

Dos Santos, V. C.; et al. 2025. Enhancing healthcare operations: a systematic literature review on approaches for hospital facility layout planning. *J. Health Organization and Management*, 39(1): 22–45.

Fahland, D.; et al. 2024. How well can large language models explain business processes? *arXiv:2401.12846*.

Fedeli, A.; and Manrique Negrin, D. A. 2024. Towards a collaborative approach for Digital Twin simulation models comprehension. In *Proc. ACM/IEEE 27th int. conf. on Model Driven Engineering Languages and Systems*, MODELS Companion '24, 660–664.

Ferrand, N.; Hassenforder, E.; and Girard, S. 2024. Engineering participation: Preparing and designing a participatory process. *Transformative Participation for Socio-Ecological Sustainability-Around the CoOPLAGE pathways*, 109–121.

Gandee, T. J.; and Giabbanelli, P. J. 2024. Combining natural language generation and graph algorithms to explain causal maps through meaningful paragraphs. In *int. conf. on Conceptual Modeling*, 359–376. Springer.

Gandee, T. J.; et al. 2024. A Visual Analytics Environment for Navigating Large Conceptual Models by Leveraging Generative Artificial Intelligence. *Mathematics*, 12(13).

Gao, M.; et al. 2023. Human-like summarization evaluation with chatgpt. *arXiv:2304.02554*.

Ghaffarzadegan, N.; Lyneis, J.; and Richardson, G. P. 2011. How small system dynamics models can help the public policy process. *System Dynamics Review*, 27(1): 22–44.

Giabbanelli, P.; et al. 2024a. Narrating Causal Graphs with Large Language Models. In *Hawaii int. conf. on System Sciences 2024 (HICSS-57)*.

Giabbanelli, P. J.; and Baniukiewicz, M. 2018. Navigating complex systems for policymaking using simple software tools. In *Advanced data analytics in health*, 21–40. Springer.

Giabbanelli, P. J.; and Vesuvala, C. X. 2023. Human factors in leveraging systems science to shape public policy for obesity: A usability study. *Information*, 14(3): 196.

Giabbanelli, P. J.; et al. 2024b. Broadening Access to Simulations for End-Users via Large Language Models: Challenges and Opportunities. In *2024 Winter Sim. Conf.*, 2535–2546. IEEE.

Giabbanelli, P. J.; et al. 2025. Promoting empathy in decision-making by turning agent-based models into stories using large-language models. *J. Simulation*.

Gierend, K.; et al. 2024. Provenance information for biomedical data and workflows: Scoping review. *J. medical Internet research*, 26: e51297.

Gray, S. A.; et al. 2013. Mental modeler: a fuzzy-logic cognitive mapping modeling tool for adaptive environmental management. In *2013 46th Hawaii int. conf. on system sciences*, 965–973. IEEE.

Haddad, E.; and Bugarin, K. 2020. Crisis Control: The Use of Simulations for Policy Decisionmaking. *Policy Brief, PB 20*, 38.

Hämäläinen, R. P.; et al. 2013. On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems. *European J. Operational Research*, 228(3): 623–634.

Hassan, S.; et al. 2024. The adoption and implementation of local government planning policy to manage hot food takeaways near schools in England: A qualitative process evaluation. *Social Science & Medicine*, 362: 117431.

He, J.; et al. 2025. Evaluating and Improving Graph to Text Generation with Large Language Models. *arXiv:2501.14497*.

Hinrichs, M.; et al. 2025. AI Integration in Mental Health Services: Examining Trends in the USA and Peoria, Illinois. In *Participatory Artificial Intelligence in Public Social Services: From Bias to Fairness in Assessing Beneficiaries*, 255–275. Springer Nature Switzerland Cham.

Huddleston, J.; et al. 2022. Design and Deployment of a Simulation Platform: Case Study of an Agent-Based Model for Youth Suicide Prevention. In *2022 Winter Sim. Conf.*, 2582–2593. IEEE.

Janssen, M.; and Helbig, N. 2018. Innovating and changing the policy-cycle: Policy-makers be prepared! *Government Information Quarterly*, 35(4): S99–S105.

Kammler, C.; et al. 2023. Towards a Social Simulation Interaction Tool for Policy Makers—A New Research Agenda to Enable Usage of More Complex Social Simulations. In *Conference of the European Social Simulation Association*, 163–176. Springer.

Keeble, M.; et al. 2024. Public acceptability of proposals to manage new takeaway food outlets near schools: cross-sectional analysis of the 2021 International Food Policy Study. *Cities & Health*, 8(6): 1094–1107.

Khademi, A.; et al. 2018. An agent-based model of healthy eating with applications to hypertension. In *Advanced Data Analytics in Health*, 43–58. Springer.

Lavin, E. A.; et al. 2018. Should we simulate mental models to assess whether they agree? In *Proc. annual simulation symposium*, 1–12.

Manellanga, R.; and David, I. 2024. Participatory and collaborative modeling of sustainable systems: A systematic review. In *Proc. ACM/IEEE 27th int. conf. on model driven engineering languages and systems*, 645–654.

Mavridis, A.; et al. 2025. Large language models for intelligent RDF knowledge graph construction: results from medical ontology mapping. *Frontiers in AI*, 8: 1546179.

Montibeller, G. 2018. Behavioral challenges in policy analysis with conflicting objectives. In *Recent advances in optimization and modeling of contemporary problems*, 85–108. INFORMS.

Mussa, O.; et al. 2024. Towards Enhancing Linked Data Retrieval in Conversational UIs Using Large Language Models. In *int. conf. on Web Information Systems Engineering*, 246–261. Springer.

Olabisi, O.; and Agrawal, A. 2024. Understanding Position Bias Effects on Fairness in Social Multi-Document Summarization. In *Proc. 11th Workshop on NLP for Similar Languages, Varieties, and Dialects*, 117–129. Mexico City, Mexico: Association for Computational Linguistics.

Omar, M.; et al. 2025. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *Int. J. Equity in Health*, 24(1): 57.

Patterson, R.; et al. 2025. Combined associations of takeaway food availability and walkability with adiposity: Cross-sectional and longitudinal analyses. *Health & Place*, 91: 103405.

Peters, U.; and Chin-Yee, B. 2025. Generalization Bias in Large Language Model Summarization of Scientific Research. arXiv:2504.00025.

Ponzo, V.; et al. 2024. Is ChatGPT an effective tool for providing dietary advice? *Nutrients*, 16(4): 469.

Rafailov, R.; et al. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proc. 37th int. conf. on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Robinson, S.; and Brooks, R. 2024. Assumptions and simplifications in discrete-event simulation modelling. *J. Simulation*, 1–18.

Sarmiento, I.; et al. 2024. Fuzzy cognitive mapping in participatory research and decision making: a practice review. *Archives of Public Health*, 82(1): 76.

Savory, B.; et al. 2025. "It does help but there's a limit...": Young people's perspectives on policies to manage hot food takeaways opening near schools. *Social Science & Medicine*, 368: 117810.

Schlicht, I. B.; et al. 2025. Do LLMs provide consistent answers to health-related questions across languages? In *European Conference on Information Retrieval*, 314–322. Springer.

Shen, J.; et al. 2024. Heart-felt narratives: Tracing empathy and narrative style in personal stories with llms. *arXiv:2405.17633*.

Shool, S.; et al. 2025. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1): 117.

Shrestha, A.; et al. 2022. Automatically explaining a model: Using deep neural networks to generate text from causal maps. In *2022 Winter Sim. Conf.*, 2629–2640. IEEE.

Singhal, K.; et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3): 943–950.

Spreng, R. N.; et al. 2009. The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *J. personality assessment*, 91(1): 62–71.

St-Aubin, B.; et al. 2023. A survey of visualization capabilities for simulation environments. In *2023 Annual Modeling and Simulation Conference*, 13–24. IEEE Computer Society.

Sun, Z.; et al. 2016. Simple or complicated agent-based models? A complicated issue. *Environmental Modelling & Software*, 86: 56–67.

Urlana, A.; et al. 2023. Controllable Text Summarization: Unraveling Challenges, Approaches, and Prospects–A Survey. *arXiv:2311.09212*.

van der Zee, D.-J. 2017. Approaches for simulation model simplification. In *2017 Winter Sim. Conf.*, 4197–4208.

Van Veen, D.; et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4): 1134–1142.

Wang, J.; et al. 2025. A recent survey on controllable text generation: A causal perspective. *Fundamental Research*, 5(3): 1194–1203.

Welivita, A.; and Pu, P. 2024. Is ChatGPT more empathetic than humans? *arXiv:2403.05572*.

Wittenborn, A. K.; and Hosseinichimeh, N. 2022. Exploring personalized psychotherapy for depression: A system dynamics approach. *Plos one*, 17(10): e0276441.

Yu, T.; et al. 2025. Diverse AI Feedback For Large Language Model Alignment. *Transactions of the Association for Computational Linguistics*, 13: 392–407.

Yu, Y.-L. 2024. Disparities by race/ethnicity and immigration status in perceived importance of and access to culturally competent health care in the United States. *J. Racial and Ethnic Health Disparities*, 11(3): 1829–1841.

Zellner, M. L.; et al. 2022. Finding the balance between simplicity and realism in participatory modeling for environmental planning. *Environmental Modelling & Software*, 157: 105481.

Zellner, M. L.; et al. 2025. Enhancing digital twin technology with community-led, science-driven participatory modeling: A case in green infrastructure planning. *Environment and Planning B: Urban Analytics and City Science*.

Zhong, M.; et al. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv:2210.07197*.