Too Open for Opinion? Embracing Open-Endedness in Large Language Models for Social Simulation

Bolei Ma[1] **Yong Cao**[2] **Indira Sen**[3] **Anna-Carolina Haensch**[1,4]
Frauke Kreuter[1,4] **Barbara Plank**[1] **Daniel Hershcovich**[5]

[1]LMU Munich & Munich Center for Machine Learning,
[2]University of Tübingen & Tübingen AI Center, [3]University of Mannheim,
[4]University of Maryland, College Park, [5]University of Copenhagen

bolei.ma@lmu.de, yong.cao@uni-tuebingen.de, dh@di.ku.dk

Abstract

Large Language Models (LLMs) are increasingly used to simulate public opinion and other social phenomena. Most current studies constrain these simulations to multiple-choice or short-answer formats for ease of scoring and comparison, but such closed designs overlook the inherently generative nature of LLMs. In this position paper, we argue that open-endedness, using free-form text that captures topics, viewpoints, and reasoning processes "in" LLMs, is essential for realistic social simulation. Drawing on decades of survey methodology research and recent advances in NLP, we argue why this open-endedness is valuable in LLM social simulations, showing how it can improve *measurement* and *design*, support *exploration* of unanticipated views, and reduce researcher-imposed *directive bias*. It also captures *expressiveness* and *individuality*, aids in *pretesting*, and ultimately enhances *methodological utility*. We call for novel practices and evaluation frameworks that leverage rather than constrain the open-ended generative diversity of LLMs, creating synergies between NLP and social science.

1 Introduction

Since NLP technologies are increasingly used in social situations, recognizing the importance of social context has become critical, and LLMs are now being applied to socially aware tasks (Hovy and Yang, 2021; Ziems et al., 2024; Yang et al., 2025). A prominent example is social simulation, where synthetic agents are used to explore collective attitudes, policy preferences, and cultural dynamics (Santurkar et al., 2023; Röttger et al., 2024; Ma et al., 2024; Anthis et al., 2025; Kozlowski and Evans, 2025, *inter alia*). A typical setup starts by prompting an LLM to "play" the role of virtual personas, conditioned on participants' profiles or past behavior data. The researcher then aggregates their responses for example, to track shifting public opinion or simulate human behaviors. Recent work highlights the promise of such simulations for generating timely, low-cost social data and for testing sociological theories at scale (Anthis et al., 2025; Wang et al., 2025a; Li et al., 2025). This potential has sparked burgeoning interest within the NLP community and beyond.[1]

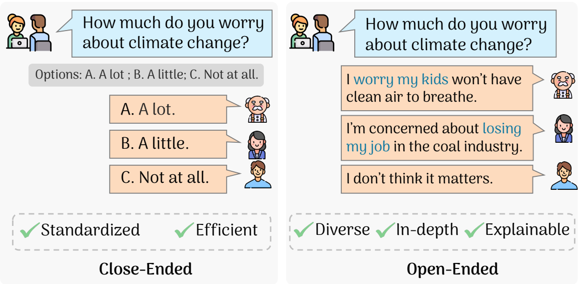

How much do you worry about climate change?
Options: A. A lot ; B. A little; C. Not at all.
A. A lot.
B. A little.
C. Not at all.
✓Standardized ✓Efficient
Close-Ended
How much do you worry about climate change?
I worry my kids won't have clean air to breathe.
I'm concerned about losing my job in the coal industry.
I don't think it matters.
✓Diverse ✓In-depth ✓Explainable
Open-Ended

Figure 1: Example of close-ended and open-ended surveying, illustrating that open-ended interaction better captures participant intentions, generates more diverse and explainable responses, and enables more accurate social simulation compared to a closed-ended one.

The majority of the most current studies usually adopt a survey-like multiple-choice or short-answer format to keep model outputs easy to score and compare (Meister et al., 2025; Balepur et al., 2025; Ni et al., 2025). While convenient, such closed forms risk collapsing nuanced opinions, steering responses toward researcher-defined categories, and masking the generative strengths of the LLMs (Lyu et al., 2024; Wang et al., 2024a,b, 2025b).

This position paper argues that social simulations should instead embrace open-endedness: the capacity of LLMs to produce free text that is as varied and unpredictable as human discourse. Open-ended questions have long been central to *sur-*

[1] See, for example, this First Workshop on Social Simulation with LLMs at COLM'25: https://sites.google.com/view/social-sims-with-llms/.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1998–2016
March 24-29, 2026 ©2026 Association for Computational Linguistics

vey methodology[2] because they surface unanticipated topics, capture minority viewpoints, and reveal reasoning processes that fixed options miss (Züll, 2016; Singer and Couper, 2017). Recent scholarship further underscores the deep parallels between NLP and survey research, noting shared methodological goals and opportunities for cross-fertilization (Kern et al., 2023; Eckman et al., 2024; Sen et al., 2025b).[3] These same properties are vital for simulations meant to approximate real social complexity. Thus, insights from decades of social science research on eliciting, coding[4], and interpreting open responses offer practical guidance for handling the variability of LLM outputs (e.g., Schonlau and Couper, 2016; Haensch et al., 2022; Gweon and Schonlau, 2023).

Position: Embracing Open-Endedness.

Embracing open-ended generation in social simulation allows researchers to capture heterogeneity in opinions, uncover minority viewpoints, and examine reasoning patterns, all of which are central to realistic social modeling.

Our paper is organized as follows: §2 introduces the concept of open-endedness in LLM social simulation. §3 reviews current practices on close-ends and limitations. §4 presents our main arguments for embracing open-endedness, drawing on survey methodology and promising use-cases. §5 offers methodological insights from social science and NLP for analyzing open-ended LLM outputs. §6 summarizes major challenges and future directions.

2 Open-Endedness in LLM Social Simulation

What is *open-endedness* or, more specifically, what are *open-ended questions* in surveys? Compared to closed-ended questions, where a fixed set of selectable response options is provided to participants, "all survey questions that do not include a set of response options are known as open-ended questions" (Züll, 2016). Their defining characteristic is that respondents are free to formulate answers in their own words, thereby introducing greater variability and richness into the collected data (Schuman and Presser, 1979; Krosnick, 1999; Bradburn et al., 2004; Krosnick and Presser, 2010; Dillman et al., 2014). Figure 1 illustrates a typical closed-ended question (left) versus an open-ended question (right).

By analogy, *open-endedness in LLMs*[5] refers to the capacity of these models to generate free-form, unconstrained text in response to a prompt, rather than being restricted to a predefined set of options. This property emerges directly from the fact they are language models, that is, they model the probability distribution over token sequences. LLMs predict the next token based on a distribution over possible continuations, without being inherently tied to a fixed response space. It is important, however, to distinguish between *open-ended generation* and *closed-ended tasks* in LLMs. Even when prompting with closed-ended questions, researchers still use models to generate text autoregressively, and extract categorical answers by mapping the generated text back to predefined labels (e.g., via string matching in Argyle et al., 2023; von der Heyde et al., 2025b) or by using token-level probabilities over constrained options (see e.g., Santurkar et al., 2023; Shu et al., 2024; Cao et al., 2025).

The crucial distinction, therefore, lies not in the generation process itself, but in the **structure of the task input design**: open-ended prompts leave the model unconstrained, whereas closed-ended prompts channel output into researcher-defined categories. For this work, we define *open-endedness* as the use of prompts that elicit unconstrained, free-form responses from LLMs without restricting outputs to a fixed set of options.

In NLP, *open-ended generation* typically refers to tasks where models are valued for creativity and diversity (Holtzman et al., 2020).[6] Our focus goes beyond: we treat open-endedness not only as artistic free play or design choice but also as a **methodological contribution for social simulation**, where the goal is to approximate the diversity of real hu-

[2] The term *survey* in survey methodology research refers to a systematic method for gathering information from a sample of a population, different from the sense used for literature review in the NLP/ML community. Survey methodology is a well-established discipline with its own principles, theories, and scientific standards in social science (Groves et al., 2009).

[3] See, also, this First Workshop on Bridging NLP and Public Opinion Research at COLM'25: `https://sites.google.com/view/nlpor2025`.

[4] In social science, *coding* refers to annotating free responses into analytic categories (He and Schonlau, 2020), distinct from the meaning of "programming" in NLP/ML.

[5] In this paper, we refer to *open-endedness*, *open-end* and *free-form generation* in LLMs as synonyms.

[6] Such as story writing (Fan et al., 2018), open-domain dialogue (Zhang et al., 2020b), or long-form question answering (Krishna et al., 2021).

man opinions and reasoning in addition to maximizing novelty or narrative flair. Here, simulation does not mean producing a single representative answer but generating multiple, potentially divergent responses that together approximate a range of plausible human perspectives (Anthis et al., 2025), and ideally including cognitively grounded reasoning (Li et al., 2025). In this sense, open-ended generation is not only a technical feature of these models but also a defining characteristic of their emerging role in simulating social processes and communicative variability.

3 Current Practice and Limitations on Close-Ends in LLM Social Simulation

Despite the generative capabilities described in §2, most LLM-based social simulations continue to favor *closed-ended* designs. From a follow-up review from the 53 studies cataloged in Anthis et al. (2025), we find that only 11 (21%) include any open-text component, and only 4 (8%) rely primarily on free-form outputs during evaluation. Similarly, in a systematic review of LLM-based social simulations, Sen et al. (2025a) find that a majority of studies use closed-ended response formats. In short, the majority of current simulations reproduce the logic of traditional survey instruments, in which models (like human respondents) are constrained to predefined response options.

This reliance on closed-ended designs for LLM-based social simulation is understandable. Multiple-choice and categorical response formats enable straightforward aggregation, facilitate quantitative comparison, and align with long-established practices in survey research (Argyle et al., 2023; Santurkar et al., 2023; Durmus et al., 2024, *inter alia*). For tasks such as measuring opinions or eliciting stated preferences from simulated agents, these formats provide scalability at low cost.

Yet these same conveniences restrict what makes LLMs distinctive. Constraining outputs to predefined labels suppresses the unexpected topics, minority viewpoints, and context-sensitive reasoning that open-ended generation can reveal (Schonlau and Couper, 2016; Gweon and Schonlau, 2023). It can also introduce *aggregation bias*, where nuanced opinions must be forced into overly broad categories (Tijdens, 2014), and risks "straightlining" or superficial responses that mask underlying attitudes. Recent critiques of multiple-choice evaluation in NLP echo these concerns, noting that fixed options reward recall over reasoning and may hide model biases or unfaithful explanations (Balepur et al., 2025). The result is a simulation that reproduces the limitations of traditional surveys rather than exploiting the richer variability of generative models.

Current practice, therefore, reveals a tension. Closed-ended tasks remain convenient, but they underuse the very property, generative openness, which distinguishes LLMs from earlier computational agents. If the goal of social simulation is to capture the diversity and complexity of human attitudes, open-ended responses should be treated not as noise to be filtered out but as a core signal to be analyzed. By drawing on the extensive literature on open-ended survey questions and LLM capabilities, we can design prompts and simulations that preserve expressive richness while supporting downstream analysis. We detail the benefits in §4.

4 Arguments for Open-Endedness

Table 1 summarizes established benefits of open-ended questions in human survey methodology and maps them to their potential implications for LLM-based social simulation. In this section, we discuss these parallels in more detail, in order to provide arguments for why open-ends are important in LLM social simulations. We begin by reviewing insights from survey research (§4.1), then extend these lessons to LLM social simulations (§4.2) to support our key position, and finally outline promising application areas (§4.3).

4.1 Lessons from Human Studies

Survey methodology has long emphasized the complementary role of open-ended questions alongside closed-ended ones. Decades of research show that open formats allow researchers to capture knowledge and attitudes in a way that structured response options cannot (e.g., Schuman and Presser, 1979; Krosnick, 1999; Bradburn et al., 2004; Groves et al., 2009; Krosnick and Presser, 2010; Züll, 2016; Albudaiwi, 2017). Drawing from these advances in survey research, we summarize the lessons and benefits in eight dimensions.

Measurement. Open-ended formats often yield more accurate knowledge measurement (Geer, 1988; Züll, 2016). Respondents' answers can better reflect what they genuinely know or recall. Although they can sometimes increase "don't know" or refusal rates, such responses themselves provide

Dimension	Human Surveys (Social Science Domain)	LLM Social Simulation (NLP Domain)
Measurement	Improves validity and reduces guessing.	Open text generation provides reasoning chains and internal consistency checks.
Exploration	Reveals unexpected or minority views.	Surfaces emergent opinions beyond predefined categories.
Design	Avoids long option lists or forced categories.	Handles large answer spaces with minimal prompt constraints.
Directive Bias	Reduces researcher-induced steering.	Allows diverse framings, lowering prompt or category bias.
Pretesting	Shows how respondents interpret questions.	Helps diagnose prompt effects and refine the simulations.
Engagement & Expressiveness	Encourages motivation and engagement with richer, more detailed articulation.	Generates varied, more expressive human-like responses beyond standardized replies.
Individuality	Captures unique, personalized phrasing.	Supports heterogeneous synthetic populations and natural variation.
Methodological Utility	Offers multi-perspective material for analysis.	Serves as "tools" and alternative framings that enrich downstream analyses.

Table 1: Benefits of open-ended questions in survey research (social science) and their implications for LLM-based social simulation (NLP), highlighting how open-ended responses enhanced corresponding tasks.

useful diagnostic information about the limits of respondents' recall or confidence (Krosnick and Presser, 2010; Albudaiwi, 2017) and the reasons for refusal (Singer and Couper, 2017).

Exploration. Open-ended questions are particularly useful when the full range of possible answers is unknown. They can surface new, unexpected, emergent, or rare perspectives that researchers might not anticipate in advance (Züll, 2016; Albudaiwi, 2017). They can help avoid priming preset answers, eliciting "first-order" thinking. (Ferrario and Stantcheva, 2022). For example, in attitude surveys, respondents may highlight concerns that fall outside predefined categories, offering insights into emergent public issues.

Design. Open-ended formats avoid excessively long lists of response options, which can overwhelm respondents when the answer space is very large, such as when classifying occupations or consumer products (Züll, 2016). Rather than presenting exhaustive lists, open-ended questions provide a direct way to capture variation more naturally, reducing the effort required in option design.

Directive Bias. Open-ended designs help reduce directive bias, which is also called "framing effect", where the way how a choice is presented can lead to selection bias (Tversky and Kahneman, 1981; Schuman and Presser, 1996; Tourangeau et al., 2000). By not steering respondents toward predefined categories, they ensure that answers reflect genuine perspectives rather than artifacts of survey design (Züll, 2016). This is particularly important in sensitive domains (Albudaiwi, 2017), where question wording and free-from response framing can shape reported attitudes. They help avoid prejudging respondents (Schuldt and Roh, 2014).

Pretesting. Open-ended questions support cognitive pretesting. They provide insights into how respondents interpret the meaning of questions and reveal the reasoning behind their choices (Lenzner et al., 2015) . Responses to the open-ended probes provide vital information on respondents' potential need for clarification and how to improve the survey questions (Singer and Couper, 2017; Neuert et al., 2021). This can be invaluable during questionnaire design, as it uncovers misunderstandings or unintended interpretations that would remain hidden in closed formats.

Engagement and Expressiveness. Open-ended questions can enhance respondent engagement, giving respondents a word (Singer and Couper, 2017). Strategically placed open-text prompts give participants space to articulate thoughts in their own words, often inviting them to share personal stories or express hesitation, reluctance, or emotional nuance that closed formats cannot capture (Albudaiwi, 2017), usually at the end of an attitude survey (Porst, 2014). This sense of involvement enriches the data and can improve quality of responses (Desai and Reimers, 2019).

Individuality. Open-ended formats capture individuality and natural variation. Each respondent's articulation is personalized with a sense of individuality, unique and sometimes creative, reflecting their particular circumstances or expressive style (Albudaiwi, 2017). This variability, though more difficult to compare systematically, is valuable: it produces data that mirrors the diversity and naturalness of real social phenomena.

Methodological Utility. Open-ended responses serve as valuable qualitative methodological resources beyond raw data contribution (Singer and Couper, 2017). Researchers can directly quote them to support findings, use them to identify relevant vocabulary, or triangulate across different perspectives on a topic. They allow studies not only to present statistical results but also to incorporate authentic "voices" to the research methodology.

4.2 Implications for LLM Social Simulations

These social science lessons carry useful implications for LLM social simulations. Open-ended responses leverage what distinguishes LLMs from previous computational approaches: their ability to generate free-form, context-sensitive text. We now draw on the implications based on findings in social science for constructing arguments for the use of open-ends if LLMs are used for social simulation, also shown in Table 1.

Measurement. Open-ended LLM responses often include justifications or contextual details, allowing researchers to assess the plausibility and internal consistency of the synthetic output. This helps determine whether the model has captured the simulated persona. For example, a conservative respondent might explain skepticism about climate change by citing economic priorities and shifting scientific claims, through an open-ended reasoning and explanation. Such reasoning, absent in closed-ended responses, mirrors the broader strategy of using a model's elicited reasoning steps as a way to validate it self's conclusions (Wei et al., 2022).

Exploration. LLMs are capable of surfacing perspectives that are not anticipated in advance (Ma et al., 2024). By using open-ended tasks, researchers can elicit emergent themes, viewpoints of hard to reach subpopulations, and unstructured arguments that mirror the unpredictability of real populations. This is especially valuable in contexts where survey research struggles, such as simulating hard-to-reach or underrepresented groups (von der Heyde et al., 2025b; Namikoshi et al., 2024). Instead of forcing all responses into predefined categories, open-ended generation allows for the discovery of attitudes and framings that would otherwise remain invisible. Such novel and complex outcomes in AI could only realized through open-ended generation (Stanley and Lehman, 2015).

Design. Traditional surveys often face design bottlenecks when the possible answer space is large or complex (Bradburn et al., 2004; Krosnick and Presser, 2010). In contrast, LLMs can easily generate responses across vast domains including political justifications, occupational identities, cultural references, without requiring the researcher to enumerate all possible categories in advance (Park et al., 2023). This not only reduces the design burden but also leads to more realistic variation in the simulated population, reflecting the breadth of how humans articulate similar underlying attitudes in many different ways.

Directive Bias. LLM outputs can be affected by how researchers prompt them (Schulhoff et al., 2025; Bardol, 2025). Closed-ended questions inevitably steer models toward predefined frames, mirroring the same risk of directive bias in human surveys. The choice of answer options and how it is presented can both affect the final results (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024; Wei et al., 2024). Open-ended prompts mitigate this risk by allowing models to generate their own framings and vocabularies. For instance, when asked about climate change or political attitudes, an LLM might emphasize scientific uncertainty, intergenerational justice, or geopolitical responsibility depending on the simulated persona, rather than being constrained to "agree/disagree" scales. This autonomy better approximates natural opinion expression and reduces researcher-induced biases in LLM outputs.

Pretesting. Just as open-ended human responses are useful for testing whether survey items are well understood, open-ended LLM responses can be used to probe how prompts are "interpreted". By inspecting generated rationales, researchers can diagnose whether a prompt unintentionally suggests an interpretation or introduces ambiguity. This supports iterative refinement of both survey instruments and simulation protocols, ensuring greater robustness. In effect, LLMs can be deployed as large-scale cognitive pretesters/advisors/pilot testers (Kim et al., 2024; Adhikari et al., 2025; Beck et al., 2025; Cui et al., 2025), helping with both real human / LLM full experiments.

Engagement and Expressiveness. Although LLMs do not face motivational constraints, open-ended generation enriches the expressive quality of simulated responses. Instead of one-dimensional choices, outputs may include narratives, hesitations, or rhetorical flourishes, which create more human-like populations. Such expressive variation is not merely decorative: it allows simulations to capture how attitudes are embedded in language practices such as sarcasm, emotional appeals, or careful hedging, adding a qualitative layer of realism that purely quantitative approaches lack.

Individuality. One of the most valuable features of open-ended LLM outputs is their ability to simulate heterogeneity. Rather than collapsing responses into standardized categories, models can generate diverse, idiosyncratic articulations of similar attitudes. This variability mirrors the individuality observed in real populations, where two people with the "same" opinion might justify it in strikingly different ways. Incorporating such micro-level differences supports the construction of synthetic agents or societies that are not only statistically representative but also qualitatively rich (Park et al., 2022, 2024).

Methodological Utility. Open-ended LLM outputs provide not only data but also analytic resources. Just as survey researchers quote respondents or analyze vocabularies, LLM researchers can use LLM-generated synthetic texts as "tools" that illustrate findings, extract salient terms to map semantic fields, or triangulate different perspectives on an issue (Liu et al., 2024; Long et al., 2024; Park et al., 2024). In this sense, open-ended simulation may not merely reproduce distributions of attitudes but also provides interpretable and quotable material, enhancing both computational and qualitative analysis of synthetic societies.

4.3 Use-Cases

Based on these benefits, we outline promising (but not exhaustive) use-cases of open-ended simulations. The detailed use-cases and implications are summarized in Appendix §A.

Populations and opinion simulation. By conditioning on demographic, ideological, or cultural profiles, LLMs can approximate subgroup-specific "silicon samples". Crucially, open-ended responses capture not only preferences but also the rationales and nuances often missed in closed-choice polls. While current LLMs may not fully replicate human behavior (Chemero, 2023; Zhang et al., 2025), studying open-ended outputs is crucial for improving realism and diversity in synthetic populations.

Democratic proxy citizens. LLMs are increasingly framed as proxy citizens in civic processes (Goñi, 2025). Such proxies offer policymakers dynamic, fine-grained insights, where the open-ended format enables evolving concerns and nuanced justifications beyond fixed-choice surveys. Despite existing gaps, analyzing free-form reasoning helps refine both model alignment and survey design.

Qualitative interviews. LLMs can emulate participants in qualitative research, generating free-form narratives in interviews and focus groups. These simulations can reveal emergent themes and discursive patterns. Current models often struggle with coherence and authenticity in extended qualitative contexts in LLM-based interviews (Kapania et al., 2025), but open-ended prompts provide a framework to study and improve these abilities.

Deliberation and debate simulations. Open-ended prompting allows LLMs to generate arguments, counterarguments, and rhetorical strategies that closed-choice setups cannot anticipate. LLMs can therefore be useful for modeling communicative dynamics, especially in multi-agent debates.

Behavior trials. In behavioral experiments, open-ended LLM responses yield richer data than fixed-choice outcomes by revealing reasoning and strategies. While LLMs cannot fully replace human participants, studying their open-ended behavior helps refine experimental design and theory testing before deployment.

Social media use-cases. LLMs can model large-scale human behavior on digital platforms. Though current models might fall short of capturing complex social dynamics (Piao et al., 2025), open-ended outputs allows agents to display diverse, context-sensitive behaviors.

Model training and evaluation. Open-ended responses are valuable not only for social science but also for advancing NLP methodology itself. We argue that social science methods can shape model alignment and evaluation. Simulated open-ended judgments yield richer signals than binary or scalar preferences, enhancing training and evaluation by revealing subtle reasoning distinctions.

5 Practical Insights from Social Science and NLP

Now, our mission is to design a concrete methodological outlook for studying the "open-ends" in social simulations. Decades of social-science practice, from meticulous qualitative coding to today's state-of-the-art language models, offer lessons for researchers who build or evaluate LLM-based simulations. We highlight how insights move in both directions: traditional social science methods such as qualitative coding and content analysis can guide the design and assessment of open-ended simulations, while advanced NLP methods including statistical machine learning workflow, LLM-driven techniques, and emerging hybrid methods can, in turn, inform social-science methodology itself, and also expand the toolkit for social simulation.

5.1 Social Science → NLP

Manual and qualitative coding. The foundational approach is manual coding supported by a well-defined codebook and reliability checks. Researchers develop conceptual categories, train multiple coders, and use double-coding or adjudication to ensure validity (Ongena and Dijkstra, 2006; Conrad et al., 2016; Singer and Couper, 2017; He and Schonlau, 2020; Neuert et al., 2021). An example of the coding scheme is shown in Appendix §B. These practices embody key social science principles, reflexivity, transparency, and intercoder reliability, that directly inform LLM simulations: synthetic personas and prompts should likewise be specified, documented, and independently reviewed to guard against hidden bias.

Quantitative and exploratory content analyses. Beyond coding, social scientists have long applied a range of traditional quantitative techniques to open-ended responses, often beginning with careful data elicitation. Web surveys, for example, exploit dynamic filtering and autocomplete to guide respondents through vast occupational or categorical taxonomies while preserving the freedom of open text (Tijdens, 2014). Once collected, responses are frequently subjected to computer-assisted qualitative data analysis (Singer and Couper, 2017) to structure the corpus and prepare it for statistical modeling. Researchers then employ exploratory text-analytic tools, such as word-frequency visualizations, keyness or distinctiveness measures, and topic modeling, to quantify patterns and highlight salient themes (Ferrario and Stantcheva, 2022; Rouder et al., 2021). These approaches illustrate how open-ended answers can be transformed into analyzable variables without losing their descriptive richness, offering practical lessons for LLM-based social simulation: careful elicitation ensures coverage of complex answer spaces, while quantitative summaries and visualizations provide scalable ways to monitor and interpret large volumes of synthetic text.

Lesson: Social Science Informs NLP.

Long-standing traditions of qualitative and quantitative analysis of the open-ends in the social sciences provide essential methodological guardrails, ensuring that when applied to NLP and LLM-based simulations, open-ended text is elicited, coded, and interpreted with rigor, transparency, and validity.

5.2 NLP → Social Science

From statistical and semi-automatic workflows to neural and Transformer-based methods. Before deep learning, social scientists turned to supervised and semi-supervised classifiers such as SVMs and random forests to scale up coding (He and Schonlau, 2021a,b). Semi-automatic workflows combine machine predictions with targeted human checks (Schonlau and Couper, 2016; Haensch et al., 2022), showing that human-in-the-loop verification can preserve rigor while improving efficiency. Recent work demonstrates the strength of transformer encoders like BERT for multi-label and multilingual survey coding (Schonlau et al., 2023; Gweon and Schonlau, 2023). This parallels the need to adapt models to particular simulated populations: just as social scientists curate training data that reflect a target electorate, simulation researchers can condition models on carefully chosen community texts or historical corpora.

LLMs for qualitative coding. A big methodological benefit is to use LLMs themselves as coders. TopicGPT (Pham et al., 2024), and other survey based coding work (e.g. Dunivin, 2025; Törnberg, 2025; von der Heyde et al., 2025a) explore zero- and few-shot coding of open-ended survey responses. These findings suggest that, with careful prompting and validation, LLMs can approach the reliability of trained human coders while providing richer reasoning traces. For social simulation, this highlights a dual role: LLMs are not only objects of study but also methodological partners for coding, auditing, and refining the very synthetic data they generate.

Hybrid strategies and human–machine collaboration. Many contemporary projects integrate human and machine strengths: analysts create the initial scheme, models propose labels, and disagreements are resolved collaboratively (Wilson et al., 2022; Haensch et al., 2022), similar to active learning in ML/NLP, where models and humans iteratively collaborate to improve labeling efficiency and quality (Settles, 2009; Zhang et al., 2022). This iterative interplay mirrors long-standing ethnographic and survey norms of triangulation, combining multiple perspectives to strengthen inference. A recent study on German opinions in LLMs (Ma et al., 2025) combined human-coded subsets with a German BERT classifier, illustrating effective human–model integration. For LLM simulations, hybrid strategies enable cyclical workflows where model outputs inform new simulations, experts provide feedback, and the model is updated or re-prompted; alternatively, experts can pre-annotate data to guide subsequent model updates.

Lesson: NLP Advances Social Science.

Breakthroughs in NLP, from traditional classifiers to transformer-based models and LLMs, actively reshape social science practices by enabling scalable coding, new modes of inference, and innovative forms of human–machine collaboration, being helpful for LLM social simulations.

6 Challenges and Future Directions

Below, we summarize three main challenges unresolved in open-ended social simulation.

Data for open-ended simulation. Progress is constrained by the lack of datasets designed specifically for open-ended social simulation. Most existing survey corpora contain only occasional free-text fields, leaving few matched human baselines for validating model output (e.g., World Value Survey Haerpfer et al., 2022). Purpose-built resources are needed in which rich demographic and contextual metadata are paired with naturally occurring open-text responses across languages and cultures. Such data would allow researchers to compare the distribution and style of synthetic opinions with real human variation, a prerequisite for assessing whether open-ended LLM populations faithfully mirror social diversity (Bender et al., 2021).

Measuring alignment in infinite responses. Open-ended social simulation makes traditional NLP metrics such as ROUGE (Lin, 2004) or BertScore (Zhang et al., 2020a) inadequate (Liu et al., 2023), since success can no longer be defined by overlap with reference texts but by authenticity through coherence, diversity, and alignment with human behavior (Park et al., 2024). Evaluation must therefore move from correctness to generative realism, supported by multi-dimensional frameworks. Core components include internal coherence with persona traits, preservation of meaning and style, and the measurement of generative diversity through metrics like the Sui Generis score (Xu et al., 2025), which captures uniqueness beyond repetitive or stereotypical outputs. It must further include statistical measures of matching the target human distribution (e.g., Huang et al., 2024), not just similarity of an individual response. Ultimately, validity also requires external correspondence with real-world data and human judgments (Elangovan et al., 2024). This triad of coherence, diversity, and empirical realism forms a structured foundation for evaluating open-ended simulations.

Ethical and societal risks. Open-ended LLM simulations raise profound ethical and epistemic concerns that extend beyond technical challenges. Open-ended generation can introduce distinct vulnerabilities compared with closed formats. As models are free to elaborate narratives arguments, they can produce plausible but unverifiable claims that later appear as emergent “public opinion”. Such unconstrained outputs make it harder to audit provenance, enabling subtle bias amplification or the spread of fabricated evidence within simulated debates. Beyond accuracy and fairness, simulations create unprecedented risks of manipulation: coordinated populations of artificial agents could be deployed to manufacture the illusion of consensus, shaping public opinion or undermining democratic processes (Goldstein et al., 2023). Finally, relying on real-world data for persona construction introduces privacy, consent, and accountability dilemmas, with blurred lines of responsibility for harmful outputs and unresolved legal issues over data and content ownership (Henderson et al., 2023).

Future directions. Addressing these challenges calls for several key directions: (i) Develop comprehensive benchmarks that pair human and synthetic open-ended responses. (ii) Move beyond surface-level similarity metrics toward multidimensional evaluation frameworks that capture coherence, diversity, empirical realism and contextual appropriateness of real human responses. (iii) Design bias-mitigation strategies that operate during generation, not only in post-hoc detection. (iv) Establish transparent, auditable, and participatory governance structures that guide data collection, model deployment, and evaluation practices.

7 Conclusion

In this position paper, we stated our position on advocating the use of open-ended questions and responses in LLM-based social simulations. Building on survey research traditions, we showed why and how methods for eliciting, coding, and analyzing free-text responses can guide both model design and evaluation. Open-ended simulation enables richer, more diverse synthetic populations and strengthens the bridge between NLP and social science. Future work should turn these links into practical frameworks that combine rigorous social science methodology with advances in LLMs.

8 Limitations

While our goal is to spark dialogue on the role of open-ended generation in LLM-based social simulation, several constraints of this position paper should be acknowledged.

Conceptual focus. Our contribution is primarily conceptual: we outline opportunities and challenges but offer no large-scale experiments or new benchmarks. The arguments rest on a synthesis of recent literature and the authors' professional experience, serving as a call for empirical studies that can validate our claims.

Non-exhaustive literature review. The review of related work is necessarily non-exhaustive, as we are a position paper, the main focus of our paper is to advocate for the awareness of this research field with examples from representative recent papers. We focus on representative papers at the intersection of NLP and survey methodology (Eckman et al., 2024; Sen et al., 2025b), extending analyses such as Anthis et al. (2025); Li et al. (2025) and referring readers to broader surveys like Ma et al. (2024), Sen et al. (2025a) and Karamolegkou et al. (2025). Our intent is to illustrate possibilities rather than to catalog the entire field.

Domain boundaries. We confine our discussion to the domain of LLM-based social simulation; open-ended methods for other NLP tasks or broader computational social science settings remain outside our scope. This deliberate focus keeps the argument anchored on how open-ended generation can uniquely enrich social simulation, which is the core position of our paper.

9 Ethical Considerations

Beyond the ethical risks regarding open-ended generation in social simulations summarized in §6, we highlight several broader ethical and epistemic concerns relevant to LLM simulation research, along with our own statement of the use of AI assistance.

Content bias of LLM responses. Recent studies of LLM-based simulations find systematic biases in model outputs, including political left-leaning tendencies and culturally WEIRD (Western, Educated, Industrialized, Rich, Democratic) framings (e.g., Argyle et al., 2023; Santurkar et al., 2023; Cao et al., 2023; Durmus et al., 2024; Ma et al., 2025; von der Heyde et al., 2025b). These representation biases risk distorting social simulations by reproducing stereotypes rather than representing diverse populations. Importantly, most of these findings stem from closed-ended setups; whether open-ended responses amplify, mitigate, or simply reshape these biases remains an open question. A recent study also shows LLM generated survey responses are more positive than human responses (Zhang et al., 2025), further questioning the validity of the LLM simulations. Future work must therefore investigate how to systematically detect and reduce representational disortions in LLM simulations, for example through bias-sensitive evaluation protocols or comparative grounding in human data.

Human-like validity bias. Human survey responses face well-known validity challenges such as satisficing, order effects, and fraudulent answering (Galesic and Bosnjak, 2009; Bless and Schwarz, 2010; Hamby and Taylor, 2016; Deaton and Stone, 2016; Kraemer et al., 2023). With the rise of automated fraud in online surveys, distinguishing genuine responses from fabricated ones has become increasingly challenging (Pinzón et al., 2024). In simulated settings, LLMs may produce fluent but strategically biased or shallowly reasoned outputs that mimic variability without genuine experience. Developing validity checks, analogous to quality control in survey methodology, will be essential to prevent mistaking coherence for authenticity in simulated data.

Epistemic risks of "silicon samples". A final concern are the broader epistemic risks of substituting simulated agents for human participants. As noted by Cummins (2025), the danger lies in building an emerging literature that reflects methodological artifacts more than substantive social phenomena. Open-ended simulations, while rich in variability, are especially prone to analytic flexibility: coding choices, prompt designs, or sampling decisions can dramatically shape findings. More specifically, whether LLMs fit within qualitative ways of "knowing" as humans, remains unclear (Kapania et al., 2025). Without transparent methodological standards, the field risks generating unreplicable insights that obscure rather than illuminate human social behavior. To avoid "silicon-only" social science, future work must couple LLM simulations with rigorous validation and clear reporting of analytic decisions.

Use of AI Assistance. The authors acknowledge the use of ChatGPT (GPT-5) exclusively to paraphrase and refine the text in the final manuscript.

Acknowledgments

We thank the members of the SODA Lab and MaiNLP Lab at LMU Munich, as well as the members of the Chair of Data Science for the Social and Economic Sciences at the University of Mannheim, for their constructive feedback. This work originated from discussions during BM and YC's visit to the University of Copenhagen, kindly hosted by DH and supported by travel grants from the Danish Data Science Academy (DDSA). BP is supported by ERC Consolidator Grant DIALECT (101043235).

References

Divya Mani Adhikari, Alexander Hartland, Ingmar Weber, and Vikram Kamath Cannanure. 2025. Exploring llms for automated generation and adaptation of questionnaires. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, CUI '25, New York, NY, USA. Association for Computing Machinery.

Dalal Albudaiwi. 2017. Survey: Open-ended questions. In *The SAGE Encyclopedia of Communication Research Methods*, volume 4, pages 1716–1717. SAGE Publications, Inc, Thousand Oaks, California.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. 2025. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2025. Plurals: A system for guiding llms via simulated social ensembles. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with LLMs? a) forced B) flawed C) fixable D) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.

Franck Bardol. 2025. Chatgpt reads your tone and responds accordingly – until it does not – emotional framing induces bias in llm outputs. *Preprint*, arXiv:2507.21083.

Jacob Beck, Stephanie Eckman, Christoph Kern, and Frauke Kreuter. 2025. Bias in the loop: How humans evaluate ai-generated suggestions. *Preprint*, arXiv:2509.08514.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Herbert Bless and Norbert Schwarz. 2010. Chapter 6 - mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. In *Advances in Experimental Social Psychology*, volume 42 of *Advances in Experimental Social Psychology*, pages 319–373. Academic Press.

Norman M. Bradburn, Seymour Sudman, and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires*, 2nd, revised edition. Jossey-Bass, San Francisco, CA.

James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using llms for market research. Working Paper 23-062, Harvard Business School Marketing Unit. Last revised: July 8, 2023.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154, Albuquerque, New Mexico. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Anthony Chemero. 2023. Llms differ from human cognition because they are not embodied. *Nature Human Behaviour*, 7(11):1828–1829.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Frederick G. Conrad, Mick P. Couper, and Joseph W. Sakshaug. 2016. Classifying open-ended reports: Factors affecting the reliability of occupation codes. *Journal of Official Statistics*, 32(1):75–92.

Ziyan Cui, Ning Li, and Huaikang Zhou. 2025. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5(8):627–634.

Jamie Cummins. 2025. The threat of analytic flexibility in using large language models to simulate human data: A call to attention. *Preprint*, arXiv:2509.13397.

Angus Deaton and Arthur A. Stone. 2016. Understanding context effects for a measure of life evaluation: how responses matter. *Oxford Economic Papers*, 68(4):861–870.

Saoirse Connor Desai and Stian Reimers. 2019. Comparing the use of open and closed questions for web-based measures of the continued-influence effect. *Behavior Research Methods*, 51(3):1426–1440.

Don A. Dillman, Jolene D. Smyth, and Leah Melani Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 4th edition. Wiley, Hoboken, NJ.

Zackary Okun Dunivin. 2025. Scaling hermeneutics: a guide to qualitative coding with llms for reflexive content analysis. *EPJ Data Science*, 14(1):28.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Stephanie Eckman, Barbara Plank, and Frauke Kreuter. 2024. Position: Insights from survey methodology can improve training data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12268–12283. PMLR.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1160, Bangkok, Thailand. Association for Computational Linguistics.

Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principals, and interventions. In *Advances in Neural Information Processing Systems*, volume 37, pages 28938–28964. Curran Associates, Inc.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Beatrice Ferrario and Stefanie Stantcheva. 2022. Eliciting people's first-order concerns: Text analysis of open-ended survey questions. *AEA Papers and Proceedings*, 112:163–169.

Stefan Feuerriegel, Abdurahman Maarouf, Dominik Bär, Dominique Geissler, Jonas Schweisthal, Nicolas Pröllochs, Claire E. Robertson, Steve Rathje, Jochen Hartmann, Saif M. Mohammad, Oded Netzer, Alexandra A. Siegel, Barbara Plank, and Jay J. Van Bavel. 2025. Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 4(2):96–111.

James Flamino, Mohammed Shahid Modi, Boleslaw K. Szymanski, Brendan Cross, and Colton Mikolajczyk. 2025. Testing the limits of large language models in debating humans. *Scientific Reports*, 15(1):13852.

Mirta Galesic and Michael Bosnjak. 2009. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2):349–360.

John G. Geer. 1988. What do open-ended questions measure? *Public Opinion Quarterly*, 52(3):365–367.

Julian Iñaki Goñi. 2025. Citizen participation and technology: lessons from the fields of deliberative democracy and science and technology studies. *Humanities and Social Sciences Communications*, 12(1):287.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *Preprint*, arXiv:2301.04246.

Robert M. Groves, Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey methodology*, 2nd edition. John Wiley & Sons, Hoboken, NJ.

Jairo F Gudiño, Umberto Grandi, and César Hidalgo. 2024. Large language models (llms) as agents for augmented democracy. *Philosophical Transactions A*, 382(2285):20240100.

Hyukjun Gweon and Matthias Schonlau. 2023. Automated classification for open-ended questions with bert. *Journal of Survey Statistics and Methodology*, 12(2):493–504.

Anna-Carolina Haensch, Bernd Weiß, Patricia Steins, Priscilla Chyrva, and Katja Bitz. 2022. The semi-automatic classification of an open-ended question on panel survey motivation and its application in attrition analysis. *Frontiers in Big Data*, 5:880554. ECollection 2022.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen, editors. 2022. *World Values Survey: Round Seven - Country-Pooled Datafile Version 5.0*. JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.

Tyler Hamby and Wyn Taylor. 2016. Survey satisficing inflates reliability and validity measures: An experimental comparison of college and amazon mechanical turk samples. *Educational and Psychological Measurement*, 76(6):912–932.

Zhoushanyue He and Matthias Schonlau. 2020. Automatic coding of text answers to open-ended questions: Should you double code the training data? *Social Science Computer Review*, 38(6):754–765.

Zhoushanyue He and Matthias Schonlau. 2021a. Coding text answers to open-ended questions: Human coders and statistical learning algorithms make similar mistakes. *methods, data, analyses*, 15(1):17.

Zhoushanyue He and Matthias Schonlau. 2021b. A model-assisted approach for finding coding errors in manual coding of open-ended questions. *Journal of Survey Statistics and Methodology*, 10(2):365–376.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *Preprint*, arXiv:2301.07543.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13441–13460, Miami, Florida, USA. Association for Computational Linguistics.

Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah E Fox. 2025. Simulacrum of stories: Examining large language models as qualitative research participants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Antonia Karamolegkou, Angana Borah, Eunjung Cho, Sagnik Ray Choudhury, Martina Galletti, Rajarshi Ghosh, Pranav Gupta, Oana Ignat, Priyanka Kargupta, Neema Kotonya, Hemank Lamba, Sun-Joo Lee, Arushi Mangla, Ishani Mondal, Deniz Nazarova, Poli Nemkova, Dina Pisarevskaya, Naquee Rizwan, Nazanin Sabri, and 13 others. 2025. Nlp for social good: A survey of challenges, opportunities, and responsible deployment. *Preprint*, arXiv:2505.22327.

Rabimba Karanjai, Boris Shor, Amanda Austin, Ryan Kennedy, Yang Lu, Lei Xu, and Weidong Shi. 2025. Synthesizing public opinions with llms: Role creation, impacts, and the future to edemorcacy. *Preprint*, arXiv:2504.00241.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback. *Preprint*, arXiv:2312.14925.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Seungone Kim, Niklas Muennighoff, Dongkwan Kim, and Alice Oh. 2025. LLM-as-an-interviewer: Beyond static testing through dynamic LLM evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26456–26493, Vienna, Austria. Association for Computational Linguistics.

Sunwoong Kim, Jongho Jeong, Jin Soo Han, and Donghyuk Shin. 2024. Llm-mirror: A generated-persona approach for survey pre-testing. *Preprint*, arXiv:2412.03162.

Austin C. Kozlowski and James Evans. 2025. Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions. *Sociological Methods & Research*, 54(3):1017–1073.

Fabienne Kraemer, Henning Silber, Bella Struminskaya, Matthias Sand, Michael Bosnjak, Joanna Koßmann, and Bernd Weiß. 2023. Satisficing response behavior across time: Assessing negative panel conditioning using an experimental design with six repetitions. *Survey Research Methods*, 17(3).

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Jon A. Krosnick. 1999. Survey research. *Annual Review of Psychology*, 50:537–567.

Jon A. Krosnick and Stanley Presser. 2010. Question and questionnaire design. In Peter V. Marsden and James D. Wright, editors, *Handbook of Survey Research*, 2 edition, pages 264–313. Emerald, Bingley, UK.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *Preprint*, arXiv:1811.07871.

Timo Lenzner, Cornelia Neuert, and Wanda Otto. 2015. *Kognitives Pretesting*. GESIS – Leibniz-Institut für Sozialwissenschaften, Mannheim.

Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, Paul Liang, Luis Alonso, and Kent Larson. 2025. Position: Simulating society requires simulating thought. *Preprint*, arXiv:2506.06958.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haijiang Liu, Qiyuan Li, Chao Gao, Yong Cao, Xiangyu Xu, Xun Wu, Daniel Hershcovich, and Jinguang Gu. 2025. Beyond demographics: Enhancing cultural value survey simulation with multi-stage personality-driven cognitive reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18406–18428, Suzhou, China. Association for Computational Linguistics.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Chenyang Lyu, Minghao Wu, and Alham Aji. 2024. Beyond probabilities: Unveiling the misalignment in evaluating large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131, Bangkok, Thailand. Association for Computational Linguistics.

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024. The potential and challenges of evaluating attitudes, opinions, and values in large language models. In *Findings of the Association for Computational Linguistics: EMNLP*

2024, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Aßenmacher. 2025. Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1785–1809, Vienna, Austria. Association for Computational Linguistics.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.

Mark-Alec Mellor and 1 others. 2023. Interview with ChatGPT. *The Irenaut*, 1(1):152–164.

Keiichi Namikoshi, Alex Filipowicz, David A. Shamma, Rumen Iliev, Candice L. Hogan, and Nikos Arechiga. 2024. Using llms to model the beliefs and preferences of targeted populations. *Preprint*, arXiv:2403.20252.

Cornelia Neuert, Katharina Meitinger, Dorothée Behr, and Matthias Schonlau. 2021. Editorial: The use of open-ended questions in surveys. *Methods, data, analyses : a journal for quantitative methods and survey methodology (mda)*, 15(1):3–6.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. A survey on large language model benchmarks. *Preprint*, arXiv:2508.15361.

Jacopo Nudo, Mario Edoardo Pandolfo, Edoardo Loru, Mattia Samory, Matteo Cinelli, and Walter Quattrociocchi. 2026. Generative exaggeration in llm social agents: Consistency, bias, and toxicity. *Online Social Networks and Media*, 51:100344.

Yfke P. Ongena and Wil Dijkstra. 2006. Methods of behavior coding of survey interviews. *Journal of Official Statistics*, 22(3):1–34.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Natalia Pinzón, Vikram Koundinya, Ryan E. Galt, William O'R. Dowling, Marcela Baukloh, Namah C. Taku-Forchu, Tracy Schohr, Leslie M. Roche, Samuel Ikendi, Mark Cooper, Lauren E. Parker, and Tapan B. Pathak. 2024. Ai-powered fraud and the erosion of online survey integrity: an analysis of 31 fraud detection strategies. *Frontiers in Research Metrics and Analytics*, Volume 9 - 2024.

Rolf Porst. 2014. *Fragebogen*, 4 edition. Studienskripten zur Soziologie. Springer VS, Wiesbaden.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Jessie Rouder, Olivia Saucier, Rachel Kinder, and Matt Jans. 2021. What to Do With All Those Open-Ended Responses? Data Visualization Techniques for Survey Researchers. *Survey Practice*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Matthias Schonlau and Mick P. Couper. 2016. Semi-automated categorization of open-ended questions. *Survey Research Methods*, 10(2):143–152.

Matthias Schonlau, Julia Weiß, and Jan Marquardt. 2023. Multi-label classification of open-ended questions with bert. In *2023 Big Data Meets Survey Science (BigSurv)*, pages 1–8.

Jonathon P. Schuldt and Sungjong Roh. 2014. Media frames and cognitive accessibility: What do "global warming" and "climate change" evoke in partisan minds? *Environmental Communication*, 8(4):529–548.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2025. The prompt report: A systematic survey of prompt engineering techniques. *Preprint*, arXiv:2406.06608.

Howard Schuman and Stanley Presser. 1979. The open and closed question. *American Sociological Review*, 44(5):692–712.

Howard Schuman and Stanley Presser. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. SAGE Publications, Inc, Thousand Oaks, CA.

Indira Sen, Marlene Lutz, Elisa Rogers, David Garcia, and Markus Strohmaier. 2025a. Missing the margins: A systematic literature review on the demographic representativeness of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24263–24289, Vienna, Austria. Association for Computational Linguistics.

Indira Sen, Bolei Ma, Georg Ahnert, Anna-Carolina Haensch, Tobias Holtdirk, Frauke Kreuter, and Markus Strohmaier. 2025b. Connecting natural language processing and survey methodology: Potentials, challenges, and open questions.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Joongi Shin, Michael A. Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding human-ai workflows for generating personas. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, page 757–781, New York, NY, USA. Association for Computing Machinery.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Eleanor Singer and Mick P. Couper. 2017. Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *methods, data, analyses*, 11(2):115–134.

Oliver Slumbers, Joel Z. Leibo, and Marco A. Janssen. 2025. Using large language models to simulate human behavioural experiments: Port of mars. *Preprint*, arXiv:2506.05555.

Kenneth O. Stanley and Joel Lehman. 2015. *Why Greatness Cannot Be Planned*, 1 edition. Springer Cham, Cham, Switzerland.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267, Miami, Florida, USA. Association for Computational Linguistics.

Kea Tijdens. 2014. Dropout rates and response times of an occupation search tree in a web survey. *Journal of Official Statistics*, 30(1):23–43.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press, Cambridge, UK.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Petter Törnberg. 2025. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6):1181–1195.

Leah von der Heyde, Anna-Carolina Haensch, Bernd Weiß, and Jessica Daikeler. 2025a. Using large language models for coding german open-ended survey responses on survey motivation. *Survey Research Methods*, 19(4):355–370.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. 2025b. Vox populi, vox ai? using large language models to estimate german vote choice. *Social Science Computer Review*, 0(0):08944393251337014.

Qian Wang, Jiaying Wu, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. 2025a. What limits llm-based human simulation: Llms or our design? *Preprint*, arXiv:2501.08579.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Rottger, and Barbara Plank. 2024a. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *First Conference on Language Modeling*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7407–7416, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xunzhi Wang, Zhuowei Zhang, Gaonan Chen, Qiongyu Li, Bitong Luo, Zhixin Han, Haotian Wang, Zhiyu Li, Hang Gao, and Mengting Hu. 2025b. UBench: Benchmarking uncertainty in large language models with multiple choice questions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8076–8107, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5598–5621, Bangkok, Thailand. Association for Computational Linguistics.

Joseph Wilson, Benjamin Pollard, John M. Aiken, Marcos D. Caballero, and H. J. Lewandowski. 2022. Classification of open-ended responses to a research-based assessment using natural language processing. *Phys. Rev. Phys. Educ. Res.*, 18:010141.

Alexander Wuttke, Matthias Aßenmacher, Christopher Klamm, Max M. Lang, Quirin Würschinger, and Frauke Kreuter. 2025. AI conversational interviewing: Transforming surveys with LLMs as adaptive interviewers. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 179–204, Albuquerque, New Mexico. Association for Computational Linguistics.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2025. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35):e2504966122.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2025. Socially aware language technologies: Perspectives and practices. *Computational Linguistics*, 51:689–703.

Simone Zhang, Janet Xu, and AJ Alvero. 2025. Generative ai meets open-ended survey responses: Research participant use of ai and homogenization. *Sociological Methods & Research*, 54(3):1197–1242.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. Can llms beat humans in debating? a dynamic multi-agent framework for competitive debate. *Preprint*, arXiv:2408.04472.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Cornelia Züll. 2016. *Open-Ended Questions*. GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany.

A Detailed Use-Cases on Open-Ends in Social Simulations

In this section, we showcase a few detailed applications as use-cases for applying the open-ends in LLM social simulations from opinion polling and civic processes to focus groups, debates, behavioral experiments, and then to improving the models themselves. These applications provide low-cost, scalable, and ethically safe testbeds for studying human-like responses and collective dynamics, opening new avenues for both exploratory and policy-relevant research.

Populations and Opinion Simulation. LLMs can be deployed as synthetic respondents to survey or poll questions, generating answers as if they were drawn from diverse human populations (Cao et al., 2025; Liu et al., 2025; Gudiño et al., 2024). Over the past years, a lot of works have been focusing on using LLMs to simulate and reflect public opinions, such as simulating the US subpopulations (e.g., Santurkar et al., 2023), German subpopulations (e.g., von der Heyde et al., 2025b) (see a comprehensive survey in Ma et al., 2024). By conditioning on demographic, ideological, or cultural backstories, models can approximate a "silicon sample" that mirrors subgroup-specific opinion distributions. For example, an LLM might be asked to respond as a working-class urban youth or a rural conservative senior, yielding distributions of attitudes across these free-form segments (Argyle et al., 2023). Applying open-ends opens possibilities for close, cost-effective, rapid opinion polling and exploratory pilot studies, enabling researchers and journalists to track emerging trends at the group level. Here, free-form answers in public opinions capture not only preferences but the rationale and nuance behind them, which are the elements that closed polls usually miss.

Democratic Proxy Citizens. Another emerging application positions LLMs as proxy citizens: digital stand-ins trained or prompted to represent individual or group preferences in civic processes. By conditioning on survey data, demographic profiles, or prior expressed views, these LLMs can generate plausible responses to unasked questions, forecast voting behavior on novel proposals, or fill participation gaps in deliberative settings. For example, Karanjai et al. (2025) propose a role-creation framework that combines demographics and personality traits to synthesize open-ended public opinion, improving alignment with real survey responses. This approach aims to provide policymakers with a more dynamic and granular proxy for public preference, augmenting traditional feedback mechanisms without the need for constant, exhaustive surveys. The open-ended format is crucial, as it allows these proxies to articulate evolving concerns or nuanced justifications that fixed-choice methods cannot anticipate, offering a richer understanding of citizen preferences.

Qualitative Interviews. LLMs can be deployed as synthetic participants or interviewers in qualitative research, such as in-depth interviews and focus group simulations (Mellor et al., 2023; Shin et al., 2024; Wuttke et al., 2025; Kim et al., 2025). By conditioning on specific social roles or demographic profiles, researchers can elicit free-form conversational narratives. In multi-agent setups, these role-conditioned LLMs can interact, producing simulated dialogues that reveal discursive patterns and emergent group themes. For instance, Slumbers et al. (2025) demonstrate how LLMs can emulate participants in the "Port of Mars" collective risk social dilemma game, exhibiting human-like cooperation, communication, and leadership in open-ended exchanges. This offers a scalable, low-cost tool for piloting interview guides and exploratory thematic analysis, enabling researchers to anticipate dynamics before committing to resource-intensive fieldwork. Though still with limitations, as shown in a recent study (Kapania et al., 2025), open-ended prompting serves as a framework to systematically explore and strengthen these abilities.

Deliberation and Debate Simulations. Open-ended prompting lets LLMs generate arguments, counterarguments, and rhetorical moves that cannot be pre-listed, creating richer deliberative spaces than closed-choice setups. Though with limitations that they cannot fully capture human nuance in debating (Zhang et al., 2024; Flamino et al., 2025), LLMs are a promising tool for modeling and exploring complex communicative interactions, especially in multi-agent scenarios (Estornell and Liu, 2024; Liang et al., 2024; Ashkinaze et al., 2025). By staging multi-agent deliberations around contested policies, scholars can analyze argumentation dynamics, persuasion strategies, and framing effects without convening live participants. For instance, Taubenfeld et al. (2024) show that while LLM agents can sustain partisan debates on salient political issues. This provides a controlled, low-cost environment to explore argumentative repertoires and potential pathways to consensus, offering researchers a tool to anticipate dynamics before engaging real-world participants.

Multi-Domain Experiments and Behavior Trials. LLMs can serve as synthetic participants in behavioral experiments where the open-ended format is crucial. LLMs can also serve as synthetic participants in controlled behavioral experiments where the open-ended format is crucial. Rather than merely making a pre-defined choice, models can provide the reasoning and strategies behind their actions in free-form text. This provides a rich layer of qualitative data, allowing researchers to analyze not just what decision was made, but the simulated cognitive process that led to it. This method is being applied across diverse fields. In psychology, for instance, LLMs generate nuanced judgments that can align with human participants (Cui et al., 2025; Feuerriegel et al., 2025). In economics, they can act as interactive agents in complex market games, with their open-ended negotiations revealing emergent strategies and behavioral biases (Brand et al., 2023; Horton, 2023). In this way, by analyzing these detailed, unscripted responses, LLM-based trials offer an open-ended space for testing theories, exploring counterfactuals, and probing the boundaries of human-like behavior before moving into real-world studies.

Social Media Use-cases. Besides augmenting surveys and behavioral experiments, another promising application of LLM simulations are large-scale simulations of human behavior on digital platforms, e.g., social media websites. Initial studies suggest that these simulations can offer a powerful tool for designing prototypes of new platforms (Park et al., 2022), e.g., designing the rules for a new subreddit, or exploratory studies of the impact of potential interventions of exiting platforms, e.g., the impact of introducing a new rule on subreddit toxicity levels. While these social media or platform simulations are not free from challenges due to the limitations of current LLM technology (Anthis et al., 2025; Nudo et al., 2026), they offer many potential benefits to traditional computational simulation techniques like agent-based models.

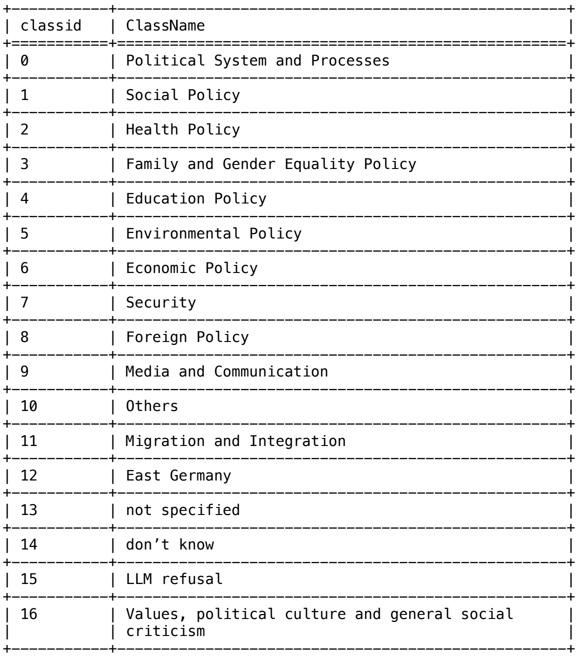
Some example simulations include *SimReddit*, a system for developing Reddit-like social media prototypes (Park et al., 2022) and the comparison of different recommendation algorithms for reducing polarization (Törnberg et al., 2023). We note that a majority of these simulations involve content-based social media platforms; while LLM-powered agents are also involved in engagement-based behaviors, e.g., liking and reporting, one of the main utilities of using LLMs instead of rule-based agents is their ability to generate 'human-like' text (ad potentially images and videos). Therefore, for this use case, it is imperative to study the open-ended simulation abilities of LLMs. Indeed there are still many open questions about how to best validate LLM-based social media simulations, given the complex, multifaceted behavior of these agents and resultant macro-outcomes of a simulation. To validate whether LLMs are good proxies of humans in social media simulations, a researcher would compare social media posts generated by LLM agents to those authored by humans. However, there are no clear evaluation standard for this comparison. One could use computational techniques (e.g., embedding-based similarity measures), qualitative techniques (human annotators being asked to differentiate generated and real posts, as in Park et al., 2022), or a combination.

In summary, social media or digital simulations are a promising use case for LLMs; but it is also a use case that primarily depends on faithful and 'human-like' open-text generations from them, making it all the more essential for NLP researchers to develop standards for evaluating open-text generations of LLMs in social simulations.

Model Training and Evaluation. The benefits of open-ended responses extend beyond social science applications and have direct implications for NLP modeling. The relationship between NLP and social science is bidirectional: while NLP offers powerful tools for advancing social good (Karamolegkou et al., 2025), insights from social science can in turn strengthen NLP modeling and evaluation practices (Eckman et al., 2024; Sen et al., 2025b). Surveys, a cornerstone of social science, are increasingly employed in NLP research to elicit both qualitative and quantitative human feedback (e.g., Argyle et al., 2023; Santurkar et al., 2023; Cao et al., 2023). Importantly, recent advances such as reinforcement learning from human feedback (RLHF), a central method for fine-tuning large language models, rely heavily on human preference data (Christiano et al., 2017; Leike et al., 2018; Ouyang et al., 2022; Kaufmann et al., 2024). In this context, open-ended responses offer a largely untapped resource. Their richness and variation provide nuanced human judgments that go beyond binary or scalar preference signals. For model training, they can expose subtle distinctions in reasoning and articulation that help align LLMs with human-like perspectives. For evaluation, they enable more human-centric assessments, offering benchmarks that capture complexity and diversity rather than reducing performance to rigid accuracy scores. In short, integrating open-ended responses into NLP pipelines can improve both the fidelity and the validity of model alignment.

B Social Science Practice: Coding Scheme

Figure 2 shows an example coding scheme of 16 categories used for coding the LLM responses in a Germany-based recent simulation study (Ma et al., 2025). This is a common practice in the social science to code the open-ended survey responses into defined categories for better measurability.



```
+-----------+--------------------------------------------------+
| classid   | ClassName                                        |
+===========+==================================================+
| 0         | Political System and Processes                   |
+-----------+--------------------------------------------------+
| 1         | Social Policy                                    |
+-----------+--------------------------------------------------+
| 2         | Health Policy                                    |
+-----------+--------------------------------------------------+
| 3         | Family and Gender Equality Policy                |
+-----------+--------------------------------------------------+
| 4         | Education Policy                                 |
+-----------+--------------------------------------------------+
| 5         | Environmental Policy                             |
+-----------+--------------------------------------------------+
| 6         | Economic Policy                                  |
+-----------+--------------------------------------------------+
| 7         | Security                                         |
+-----------+--------------------------------------------------+
| 8         | Foreign Policy                                   |
+-----------+--------------------------------------------------+
| 9         | Media and Communication                          |
+-----------+--------------------------------------------------+
| 10        | Others                                           |
+-----------+--------------------------------------------------+
| 11        | Migration and Integration                        |
+-----------+--------------------------------------------------+
| 12        | East Germany                                     |
+-----------+--------------------------------------------------+
| 13        | not specified                                    |
+-----------+--------------------------------------------------+
| 14        | don’t know                                       |
+-----------+--------------------------------------------------+
| 15        | LLM refusal                                      |
+-----------+--------------------------------------------------+
| 16        | Values, political culture and general social     |
|           | criticism                                        |
+-----------+--------------------------------------------------+
```

Figure 2: An example coding scheme modified for coding open-ended LLM responses to a question “In your opinion, what is the most important question facing Germany today?”. The human-coders (or LLM-Coders) are instructed to categorize the open-ended responses into the listed classes.