Improving Interpersonal Communication by Simulating Audiences with Large Language Models

Ryan Liu (ryanliu@princeton.edu)
Department of Computer Science, Princeton, NJ 08540 USA

Howard Yen (hyen@princeton.edu)
Department of Computer Science, Princeton, NJ 08540 USA

Raja Marjieh (raja.marjieh@princeton.edu)
Department of Psychology, Princeton, NJ 08540 USA

Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science, Princeton, NJ 08540 USA

Ranjay Krishna (ranjay@cs.washington.edu)
Allen School of Computer Science, Seattle, WA 98195 USA

Abstract

How do we communicate with others to achieve our goals? We use our prior experience or advice from others, or construct a candidate utterance by predicting how it will be received. However, our experiences are limited and biased, and reasoning about potential outcomes can be difficult and cognitively challenging. In this paper, we explore how we can leverage Large Language Model (LLM) simulations to help us communicate better. Based on ideas from cognitive science such as the Rational Speech Act model, we propose the Explore-Generate-Simulate (EGS) framework, which takes as input any scenario where an individual is communicating to an audience with a goal they want to achieve. EGS (1) *explores* the solution space by producing a diverse set of advice relevant to the scenario, (2) *generates* communication candidates conditioned on subsets of the advice, and (3) *simulates* the reactions from various audiences to determine both the best candidate and advice to use. We evaluate this framework on eight scenarios spanning a range of interpersonal communication settings. For each scenario, we collect a dataset of human evaluations across candidates and baselines, and show that our framework's chosen candidate is significantly preferred over popular generation mechanisms for LLMs. Finally, we demonstrate the generality of our framework by applying it to real-world scenarios described by users on web forums.

Keywords: Large language models; Goal-oriented communication; Large language model simulation

Introduction

We communicate with others in order to achieve our goals: to make friends, to accomplish tasks, or simply to convey our intentions (Grice, 1975; Sperber & Wilson, 1986). However, it can be hard to find the right words to achieve those goals. Consider a scenario where you are trying to get a discount on an item by haggling with its vendor. There are many strategies that you could use, including complimenting the item, offering to buy multiple items for a discount, or even describing your financial situation and asking them to take pity. With so many potential options, it is difficult to correctly decide which strategy to choose. This problem is not confined to bargaining: everyday communication requires us to make choices about what approach to take, whether making friends, impressing others, or navigating romantic conflicts.

Given a communication scenario, how do we decide which strategies to employ? Often, we rely on heuristics such as our prior experience (Schacter, Addis, & Buckner, 2007) or on advice we receive from others (Yaniv, 2004). When we have more time to make careful decisions, we may even play out possible candidates in our minds, simulating the reaction of an imaginary listener and using their imagined reaction to guide our choice (Atance & O'Neill, 2001). This idea is formalized in the Rational Speech Act (RSA) model (Goodman & Frank, 2016), which explains people's communication choices in terms of speakers simulating listeners as rational interpreters of possible candidate utterances. However, both our experiences and the advice of others are biased by the information we are exposed to, making our heuristics and simulations imperfect and resulting in suboptimal communication outcomes (Gilbert & Wilson, 2007). Moreover, reasoning about others' potential reactions can be time-consuming and cognitively challenging (Gilbert, Pelham, & Krull, 1988).

Inspired by the newfound capacity of large language models (LLMs) to simulate agents (Park et al., 2023), we propose the Explore-Generate-Simulate (EGS) framework, which supports people in exploring communication strategies and developing message candidates while offloading the cognitively challenging simulation of audience reactions. More precisely, given an arbitrary communication scenario, EGS first *explores* the space of possible responses by using an LLM to produce both normal and creatively unorthodox advice relevant to the scenario. Next, it *generates* communication candidates by conditioning an LLM on various subsets of the advice. Finally, it *simulates* the reception of each candidate by having the LLM take on the perspectives of key audiences. Using these simulations, we can estimate which candidates and advice are best suited for achieving the communicator's goal.

To evaluate this framework, we construct eight diverse scenarios that span a variety of communication modalities, relationships, and settings (see Table 1). For each, we use EGS to generate candidate messages, and collect human judg-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

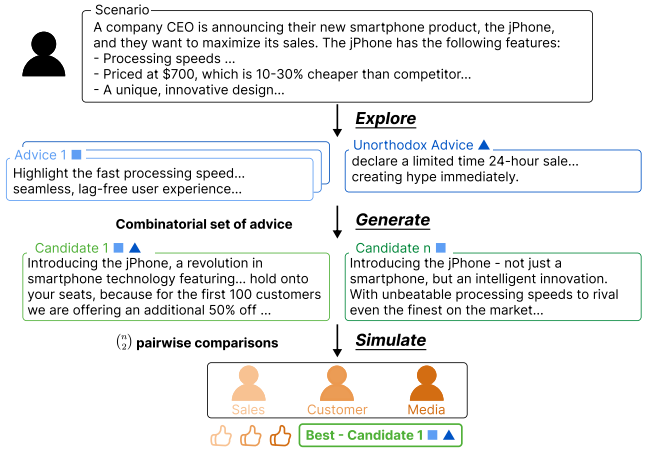

Figure 1: Given a scenario and goal, EGS generates the best candidate message by simulating stakeholders using an LLM. It *explores* pieces of advice that might help, *generates* candidates conditioned on subsets of advice, and *simulates* audience members who evaluate the various candidates.

ments on the effectiveness of each. We also compare how EGS's final recommended candidate performs against non-simulation baselines, including GPT-4 zero-shot and Chain-of-Thought (CoT). We find that EGS significantly outperforms these baselines, and that the *Simulate* step achieves significant agreement with crowdsourced human ratings. Separately, we use real world scenarios drawn from the Stanford Human Preferences (SHP) dataset (Ethayarajh, Choi, & Swayamdipta, 2022) to further analyze the agreement between humans and simulated audiences in a wider domain, where our approach also convincingly outperforms baselines.

Table 1: Constructed scenarios and communicative goals.

Scenarios
Airline minimizing negative public opinion after plane crash.
CEO giving a product announcement to maximize sales.
Customer negotiating the price of a product with an artisan.
Barista trying to maximize their tip with a customer.
Office worker getting closer to a coworker by sharing secrets.
Young man trying to get matches on a dating app.
Teenager trying to give a white lie to his date about her outfit.
Wife trying to confront husband without starting a fight.

Background

Our work uses scenarios motivated by the literature on **interpersonal communication**. Our framework tries to capture **mental simulation** with LLM **agent simulations**. We introduce these ideas in turn.

Interpersonal communication. Classical formal models of communication view cooperative communication as information transfer between speaker and listener (de Saussure, 1916; Lewis, 1969; Shannon, 1948), with a focus on aligning the true state between agents as the goal of communication (Stalnaker, 1978). The linguist H. Paul Grice identified a set of maxims surrounding the problem of how a cooperative speaker should choose what to say, such as truthfulness or relevance (Grice, 1957, 1975, 1989). The Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016) draws on both these perspectives, modeling informative speakers as aiming to reduce the listener's uncertainty over the true world state, assuming listeners make rational inferences from the utterances they hear.

Mental simulation. The act of projecting oneself into the future to pre-experience an event is formalized in cognitive science as "episodic future thinking" (Atance & O'Neill, 2001). At the neuropsychological level, brain regions associated with memory are similarly engaged when people imagine future experiences (Schacter et al., 2007). Szpunar (2010) uncovers a close relation between episodic future thought and the ability to remember personal episodes from one's past. In cognitive psychology, (Klein, Robertson, & Delton, 2010) find evidence that one goal of long-term memory is to store information about the past to plan for the future. Thus, if we view LLMs as artifacts that encode an aggregation of personal experiences across a large subset of human society, LLMs may also have the capacity to simulate epsiodic future thought. Furthermore, as LLMs are theoretically capable of taking experiences and inferring and representing properties of an agent likely to have had those experiences (Andreas, 2022), they may be able to simulate episodic future thought from the perspective of an agent with particular experiences, leading to the possibility of simulating audiences.

Agent simulations. LLM simulations are seen as an opportunity to expand computational social science research (Ziems et al., 2023). Although rule-based simulations have traditionally been used to study social phenomena, they are limited in expressivity (Schelling, 1971; Easley & Kleinberg, 2010). LLMs can potentially simulate more complex interactions that are harder to codify. They can also be explicitly conditioned to simulate individuals with goals and objectives (Jones & Steinhardt, 2022; Koralus & Wang-Maścianica, 2023; Liu & Shah, 2023). This was recently used to simulate social computing systems (Park et al., 2022), roll out their members' interactions (Park et al., 2023), generate public opinion (Chu, Andreas, Ansolabehere, & Roy, 2023), and even produce individualized subjective experience descriptions (Argyle et al., 2023). However, there are uncertainties in using LLM simulations due to their unpredictability (Reiss, 2023; Salinas & Morstatter, 2024). LLMs also exhibit higher homogeneity of opinions than humans (Argyle et al., 2023; Santurkar et al., 2023), and combining LLMs with human samples is essential to avoid algorithmic monoculture representing only a limited set of perspectives (Kleinberg & Raghavan, 2021; Bommasani et al., 2021).

A Framework for Supporting Communication

We focus on the following setting: An AI system is given as input an arbitrary scenario where an individual is communicating to an audience with a goal they want to achieve. As output, it suggests a candidate message and set of advice to help the individual achieve their goal. In this setting, standard methods for generating messages using LLMs face three challenges. First, LLMs often lack diversity in their wording and approach. Second, the communicator might lack perspectives from important audiences in order to accurately evaluate one message candidate against another. Third, as more potential candidates are generated, the user can get easily overwhelmed when deciding between the options, requiring a more scalable method of making judgements between message candidates. The Explore-Generate-Simulate framework is designed to address these challenges simultaneously.
An example input may be

> *A company CEO is announcing their new smartphone product, the jPhone, and they want to maximize its sales. The jPhone has the following features [. . .]*
> *They are about to give a 30-second presentation about the jPhone, broadcasted to major television channels.*

Here, the communicator is the CEO, and their communicative goal is to *maximize the sales of the jPhone*. In the **Explore** step of the framework, a LLM generates a diverse set of advice for the scenario. One example could be

> *Highlight the fast processing speed and seamless, lag-free user experience.*

In the **Generate** step, EGS prompts a LLM to generate candidates for the communication based on subsets of the advice. For the advice above, an example candidate is

> *Introducing the jPhone. With unbeatable processing speeds to rival even the finest on the market. . .*

In the **Simulate** step, EGS asks a LLM to generate a list of stakeholder audience profiles, each with a unique description and perspective. For the above, an example stakeholder and their perspective is

> ***Media Outlets:** Your job is to listen to the CEO's presentation, understand the key features and selling points of the phone, and relay information to the public. . .*

The LLM then takes the role of each simulated audience and evaluates each candidate based on the likelihood and magnitude to which the candidate achieves the communication goal.

The EGS framework is modular, allowing each step to be implemented flexibly based on the specific scenario. We now go into each step with more detail.

Explore

The purpose of the *Explore* step is to expand the space of possible candiate generations. This step generates a list of distinct pieces of advice to later condition the candidate generation upon. We follow existing literature in Social Psychology, which finds that people recall useful advice (Yaniv, 2004) or prior experiences (Schacter et al., 2007) when considering their next action. Similarly, *Explore* generates relevant pieces of advice that will be useful for the next stage.

EGS also prompts LLMs to generate "unorthodox but potentially helpful" advice to increase diversity. For instance, in the example above, GPT-4 generates the unorthodox advice:

> *declare a limited time 24-hour sale where the first 100 customers get the phone at an additional 50% off, creating hype and urgency to buy immediately.*

Qualitatively, we find unorthodox advice generated by GPT-4 is clever and creative, while quantitatively it improves generated candidates in 4 of 8 scenarios (see Supplement B.8).[1]

Generate

The *Generate* step seeks to create reasonable candidates for communication guided by advice from the Explore step. EGS forms combinatorial sets of the generated advice up to k pieces at once, where k is an adjustable hyperparameter, and each subset is used to generate multiple candidate messages.

Following Park et al. (2023), we use the "inner voice" of the communicator to condition generated candidates on their assigned advice set by including a memory in context, making the LLM more likely to treat the statement as a directive:

> [Scenario Description]
> *You remember a few pieces of advice: [. . .] You decide to focus on using these pieces of advice during your [. . .]*

Conditioning on a combinatorial spread of advice further expands the explored solution space. This allows each candidate to incorporate orthogonal advice concepts, which we find leads to better performance through an ablation (see Supplement B.10). We also conduct a preliminary investigation into generating candidates conditioned on specific audiences in addition to advice, and find that they are not preferred over those generated with only advice (see Supplement B.12).

Simulate

In the *Simulate* step, EGS first prompts an LLM to generate a list of audiences that directly influence the communication goal, and construct a profile description for each. Next, EGS simulates the reactions of these audiences to each candidate message, and aggregates the results to estimate which candidate message best achieves the communicator's goal.

For each audience, EGS prompts the LLM to construct 1) a description of the scenario and reception of a communicated message from their point of view, 2) a weighted score for the audience's relative importance, and 3) the appropriate question regarding how their reaction to a candidate message directly influences the communicator's goal. For example, for the *media outlets* audience, the question generated was

[1] Supplement is available at this link: https://osf.io/5an29/?view_only=de721c79a50b4259af7ec759bc823f0e

In which scenario would you be more likely to give more media coverage and promotion towards the jPhone?

We aggregate audience evaluations using the generated weights and compute the simulated best candidate and advice via weighted sum. For prompts used for audience generation and examples, see Supplement A.5.

Since candidates can be close in quality, LLMs can lack granularity when giving ratings to individual candidates (Qin et al., 2023), so we ask LLM simulated audiences to use pairwise comparisons to express more detailed preferences. In the prompt, we provide the LLM with a simulated audience profile and two scenarios, one representing each candidate in the pairwise comparison. We then ask the LLM to reason about which is better before providing an outcome $o \in$ {"prefer scenario 1", "prefer scenario 2", "tie"}. Once we have outcomes for each audience a, we aggregate weighted scores across audiences to get simulated best candidate c^*.

$$c^* = \max_c \sum_{c' \neq c} \sum_a w_a \cdot compare(a, c, c') \quad (1)$$

$$compare(a, c, c') = \begin{cases} 1 & \text{if "audience } a \text{ prefers } c\text{"} \\ 0.5 & \text{if "tie"} \\ 0 & \text{if "audience } a \text{ prefers } c'\text{"} \end{cases} \quad (2)$$

Human Evaluations

We collected ratings from human participants to evaluate how EGS's recommended advice and candidate perform versus other methods. We also use these ratings to measure agreement between humans and *Simulated* audience judgments.

Data Collection

For each scenario, we collect human ratings of the quality of all candidates from *Generate*, as well as two baselines (GPT-4 zero-shot and CoT). In our experiments, we use three normal and one unorthodox piece of advice during *Explore*. We limited advice sets to ≤ 2 pieces of advice, yielding 10 sets, and three candidates per set to create 30 candidates per scenario.

Baselines. We used two baselines, GPT-4 zero-shot and chain-of-thought (Wei et al., 2022, CoT). In zero-shot, we provide the scenario description, communicator profile, and communication goal and prompt GPT-4 to generate an utterance. In CoT, we ask GPT-4 to reason about the scenario before responding with an utterance (prompts in Supplement A.7.2). We also conduct an ablation for *Explore*, where we prompt the LLM to generate advice that is encouraging or irrelevant instead. For both, we explore three pieces of that type of advice, create six advice sets of size ≤ 2, and generate 18 ablation candidates per scenario.

Ratings. Human ratings were on a 0-10 Likert scale, with a scale from (0) highly negative to (10) highly positive results for achieving the communicator's goal. In the example,

Table 2: The best candidate message selected by EGS outperforms GPT-4 zero-shot in human ratings across all constructed scenarios, and outperforms GPT-4 with CoT in five scenarios and a subset of a sixth scenario.

Scenario	GPT-4 zero-shot	Chain-of-Thought	EGS (ours)
Plane Crash	6.83 ± 0.21**	5.98 ± 0.20***	**7.95** ± 0.30
Product Launch	5.73 ± 0.26**	5.95 ± 0.25*	**7.05** ± 0.44
Bargaining	4.68 ± 0.29	**5.98** ± 0.26	5.85 ± 0.72
Barista	5.58 ± 0.29	**5.78** ± 0.26	5.40 ± 0.47
Sharing Secrets	3.67 ± 0.21***	4.17 ± 0.26**	**5.55** ± 0.46
Dating App	5.42 ± 0.30	**6.48** ± 0.26	5.05 ± 0.52
White Lie During Date	6.12 ± 0.26	6.02 ± 0.27	**6.70** ± 0.54
Marriage Argument	6.78 ± 0.28*	6.70 ± 0.28*	**7.80** ± 0.39
Average	5.60 ± 0.10***	5.88 ± 0.10**	**6.42** ± 0.19

Note: *, **, and *** denote $p < 0.05$, 0.01, and 0.001 when compared to EGS. Errors are standard errors of the mean.

"If you were considering getting a new phone, how likely are you to buy a jPhone?"

(0)*"Not at all"*...(5)*"Somewhat likely"*...(10)*"Definitely"*

Participants. Our human ratings comprised 12180 human judgments from $N = 652$ UK participants crowdsourced via Prolific. Participants provided informed consent prior to participation in accordance with an approved institutional review board protocol, and were paid 12 USD per hour. We collected 20 judgments per candidate and baseline, and each participant provided ≤ 20 judgments. This yielded an average inter-rater reliability of $r = .82$. More details in Supplement A.8.

EGS outperforms GPT-4 Zero-shot and CoT

Averaging across all scenarios, the average human ratings of EGS outperformed the GPT-4 zero-shot baseline by 0.82 (14.6%) and CoT by 0.54 (9.2%), both statistically significant at $\alpha = 0.01$ using bootstrapping with 10000 samples. Separating by scenario, EGS outperformed GPT-4 zero-shot in all scenarios, and CoT in five scenarios (Table 2), of which four each are statistically significant at $\alpha = 0.05$.

All outputs from EGS surpassed a mean score of 5, indicating that they all had a positive impact on the communicator's goal, whereas this was not the case for either baseline. In the Bargaining scenario, we find a large discrepancy between human and GPT-4 preferences on the unorthodox advice (see Supplement B.1 for details). After reducing the *Explore* space by removing the unorthodox advice, EGS outperforms CoT by a large margin in this scenario (5.85 $\rightarrow$ 6.60 vs. 5.98).

We also find that the *Explore* step outperforms ablations that generate encouraging or irrelevant advice, with significant improvements across 6/8 scenarios (Supplement B.7).

Significant agreement between humans & GPT-4

Multilevel model across scenarios. Using a multilevel model, we analyze the agreement between GPT-4 and human raters, assigning scenario as a random effect. We measured if candidates preferred in pairwise comparisons by EGS had

Table 3: We find significant agreement across GPT-4 and human ratings in five scenarios and a subset of a sixth scenario. Preferred / less preferred values are mean scores across LLM comparisons, with standard errors of the mean. Agreement denotes percentage agreement across pairs.

Scenario	Preferred	Less Preferred	Agreement
Plane Crash	**6.19 ± 0.03*****	5.86 ± 0.03	0.63
Product Launch	**6.20 ± 0.03*****	5.87 ± 0.03	0.67
Bargaining	5.90 ± 0.03	**5.99 ± 0.02***	0.53
Bargaining (-unorthodox advice)	**6.35 ± 0.04*****	6.06 ± 0.03	0.69
Barista	**4.66 ± 0.08*****	3.53 ± 0.09	0.64
Sharing Secrets	**5.72 ± 0.03*****	4.99 ± 0.04	0.78
Dating App	5.24 ± 0.03	**5.44 ± 0.03*****	0.41
White Lie During Date	6.70 ± 0.03	**6.81 ± 0.03****	0.43
Marriage Argument	**6.34 ± 0.04*****	6.01 ± 0.03	0.65

Note: *, **, and *** denote $p < 0.05, 0.01$ and 0.001.

higher mean scores from human raters. We find a significant fixed effect for pairwise judgements on the score provided by human raters $(\text{coef} = 0.427, p = 0.041)$, demonstrating significant agreement between simulated and human scores across scenarios. This effect differed across scenarios, indicating the multilevel model's appropriateness in taking into account the hierarchical nature of our data.

GPT-4 comparisons vs. human ratings per scenario. For each scenario, we conducted a paired samples t-test across the preferred and less preferred candidates of the pairwise comparisons, and find that the preferred have significantly higher scores in 5/8 scenarios with $\alpha = 0.001$ (see Table 3).

While ratings allows us to perform statistical tests, they are more strongly influenced by easier comparisons that have a large disparity in scores. Thus, we follow with a percentage agreement analysis where each pair is weighted equally.

Percentage agreement within individual scenarios. For each pair of candidates, we aggregate pairwise comparisons made by audiences using a weighted sum, and compare the outcome with the mean human scores of each candidate to see if they match. Tied comparisons are labeled as a half-match. We divide matching pairs by the total pairs to obtain percentage agreement. For a mathematical formulation and justification of our metric, please refer to Supplement B.4. In five scenarios and the modified bargaining results, we find agreement > 0.6 between human raters and GPT-4 (Table 3).

Broader internet user simulation

We further evaluate EGS's audience simulation on a broader space of interaction using the Stanford Human Preferences (SHP) dataset (Ethayarajh et al., 2022). SHP contains 385K human preferences over responses to online forum posts across a wide variety of subject areas from cooking to legal advice, making it a robust test bed for the simulation of different audiences. Each SHP entry contains a forum post, two comments from the discussion, and the number of upvotes each comment received. We provide examples of questions and scenarios in Supplement C.

Table 4: EGS conditioned on different audience prompts outperforms GPT-4 w/ CoT on reddit user preferences. Legaladvice and askculinary are more casual, where a fun-seeking profile performed better, while default audience was more accurate on serious forums such as asksocialscience.

Domain	CoT	EGS Redditor (Default)	EGS Redditor (Funny)
legaladvice	71.0	70.0	**76.0**
askculinary	59.0	60.0	**70.5**
askhr	72.0	**76.0**	72.5
eli5	74.0	**76.0**	71.5
asksocialscience	77.8	**79.4**	61.9

We evaluate three methods on their accuracy of predicting the comment with more upvotes. First, we use a CoT baseline, which prompts the LLM to reason about the post before predicting. Next, we use two versions of *Simulate*, each with an audience that we define beforehand. EGS Redditor simulation (Default) takes the perspective of a Redditor browsing the forum, and prompts the LLM to reason about which comment it is more likely to upvote; EGS Redditor simulation (Funny) additionally specifies that the Redditor is more likely to upvote funny and entertaining comments. Prompts and output examples can be found in Supplement C.1. Following the creators of the dataset, we filter SHP by a ratio threshold of 3, ensuring that the more preferred comment is nontrivially preferred in each pair. To reduce the cost of API access, we selected 5 subreddits and randomly sample 100 pairs from each, resulting in a total of 500 evaluations.

We observe that EGS Redditor simulation is equal or better than the CoT baseline (Table 4), suggesting that LLMs are able to make decisions more aligned with real users when explicitly prompted to simulate them. Directing the LLM to look for funny/entertaining comments boosted performance on more casual forums such as cooking and legal advice. In domains such as as socialscience where strict rules are enforced on informativeness and sincerity, the demographic of viewers matched Redditor Simulation (Default) more, and performance increased accordingly. We perform a more cohesive investigation into redditor personalities and how they affect performance in Supplement C.2. Our results further validate that the *Simulate* method generalizes to diverse domains of discussion, and demonstrate that performance can be further boosted with a better understanding of audiences.

Discussion

Listening to advice and considering potential audiences are natural steps in crafting effective communication. In this paper we present results that these steps can be partially automated using LLMs. Our results suggest that the Explore-Generate-Simulate process is an effective way to prompt LLMs to produce useful messages across a range of settings. Next, we consider some implications of these results.

Computation load & Scalable episodic future thinking A key contribution of EGS is as a scalable alternative to episodic future thinking. In our experiments, each simulated audience in each scenario performed 1305 pairwise comparisons between candidate messages with detailed reasoning. This took between 2-4 hours with one API key at 10000 tokens/min, or one simulated comparison every 6-11s. Since then, API usage limits have risen to 30 million tokens/min, allowing a simulated comparison to be conducted in less than 0.01 seconds. Additionally, human simulations of future events are limited by the linear stream of consciousness over time, but EGS can be parallelized to achieve even faster speeds. This, in addition to the significant agreement between simulated audiences and humans, makes EGS a viable alternative when preparing for communication events that would normally require extensive episodic future thinking.

Reasoning about the past Aside from optimizing communication in the present, EGS can also be applied to perform counterfactual reasoning (CFR) over past communication. Given a past scenario and its outcome, CFR concerns whether an alteration to the antecedent of the counterfactual affects the outcome (Pearl, 2000).

Given a past communication setting, we can use *Explore* and *Generate* to create a diverse list of alternatives to the antecedent, and then *Simulate* outcomes when we replace the antecedent with each alternative. In this application, we can also include the communication utterance used and the actual outcome as a known ground truth in context for the LLM. Using simulated pairwise comparison results, we can reason about which utterances could have potentially been used to reach a better outcome, or if any pieces of advice were responsible for a particular end result. Though these causal effects may not be directly transferrable to the real world due to simulation inaccuracies, they provide testable hypotheses that can be directly implemented in behavioral experiments.

LLMs as shared cultural experience Our system also makes it possible to share information from LLMs' training data that surpasses individual, cultural, or geographical barriers. For example, a barista may not pay attention to a five star customer review halfway across the world about a pleasant experience, but LLMs may have the capability to connect and synthesize a wide range of accounts into the responses they generate. Similarly, a person may not have a habit of reading online forums for relationship advice, but LLMs may be able to take inspiration from these sources to provide meaningful insights. In constructing this framework, we also hope to connect people with communication strategies that might not be available to them, not just to improve their communication but also potentially helping them grow as communicators.

Granularity of simulation A key question that remains about audience simulation is whether there is a "sweet spot" for the level of detail a simulated audience should have. When we ask the LLM to generate stakeholder audiences, we do not specify the level of detail at which they are generated. Consequently, we noticed that many simulated comparisons included a disclaimer about how the audience's preference depends on various details, suggesting that a more detailed description of generated audiences may improve performance. However, a higher level of detail could also result in less representative audiences or less accurate simulations.

This effect is not just limited to audiences but also to the scenario. Details of varying importance in the scenario can be included or omitted in the input to EGS. For instance, the plane crash scenario might also include a recent crash that happened within the same company. We observe varying levels of hallucination in generated candidates (see Supplement B.15 for an analysis), with many partially caused by a lack of information provided in the scenario — the model is forced to hallucinate in order to follow *Explored* advice. Ultimately, there is a trade-off between accurately replicating the scenario and making the described scenario easier for the LLM to reason about or simulate.

Envisioned uses and risks Though EGS can be used to assist in any communication, we believe that EGS can particularly assist when (1) the audiences are new or unfamiliar and it is important to consider their point of view, (2) the search space for the potential message is large, and (3) the audience is composed of a combination of people with individual traits and differences to cater towards. EGS can also potentially assist individuals who experience difficulties communicating with others by providing multiple concrete examples of plausible messages. Furthermore, EGS could also allow ideas to be shared in ways that are more acceptable in order to foster communication between groups that are polarized or divided.

Despite our focus on positive use cases, we acknowledge that EGS is dual-use and can be used to optimize communications detrimental to society. While EGS can simulate readers to improve email clarity, it can also increase the success of phishing or for creating propaganda. While LLMs are safety-tuned to avoid these behaviors, this does not avoid usage with base models. Safety-tuned LLMs may also exhibit reward-hacking behaviors by selecting utterances that superficially improve communication through deceit or manipulation. Nonetheless, as models improve in recognizing and refusing queries with malicious intent (Touvron et al., 2023; Huang, Gupta, Xia, Li, & Chen, 2023), safety concerns surrounding general use of EGS will also improve.

Limitations and future work While we compare against LLM baselines, another natural baseline would be to compare against human responses. Additionally, while we recruited participants from the UK, future work could assess generalizability to other populations. Lastly, LLMs may also reflect historical biases (Weidinger et al., 2021). While simulating diverse stakeholders may help, more work remains to be done to simulate such profiles accurately.

References

Andreas, J. (2022). Language models as agent models. *arXiv preprint arXiv:2212.01681*.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Atance, C. M., & O'Neill, D. K. (2001). Episodic future thinking. *Trends in cognitive sciences*, *5*(12), 533–539.

Bai, Y., et al. (2022). *Training a helpful and harmless assistant with reinforcement learning from human feedback.*

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, W. (1910). Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology*, *3*(3), 296–322.

Chu, E., Andreas, J., Ansolabehere, S., & Roy, D. (2023). Language models trained on media diets can predict public opinion. *arXiv preprint arXiv:2303.16779*.

de Saussure, F. (1916). *Course in general linguistics* (P. Meisel & H. Saussy, Eds.). Columbia University Press.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Ethayarajh, K., Choi, Y., & Swayamdipta, S. (2022, Jul). Understanding dataset difficulty with $\mathcal{V}$-usable information. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (pp. 5988–6008).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *334*(6084), 998–998.

Gates, V., Griffiths, T. L., & Dragan, A. D. (2020). How to be helpful to multiple people at once. *Cognitive science*, *44*(6), e12841.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of personality and social psychology*, *54*(5), 733.

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science*, *317*(5843), 1351–1354.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Grice, H. P. (1957). Meaning. *Philosophical Review*, *66*(3), 377–388.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.

Huang, Y., Gupta, S., Xia, M., Li, K., & Chen, D. (2023). *Catastrophic jailbreak of open-source llms via exploiting generation.*

Jain, S., Ma, X., Deoras, A., & Xiang, B. (2023). *Self-consistency for open-ended generations.*

Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, *35*, 11785–11799.

Klein, S. B., Robertson, T. E., & Delton, A. W. (2010). Facing the future: Memory as an evolved system for planning future acts. *Memory & cognition*, *38*, 13–22.

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, *118*(22), e2018340118.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large language models are zero-shot reasoners.*

Koralus, P., & Wang-Maścianica, V. (2023). Humans in humans out: On gpt converging toward common sense in both success and failure. *arXiv preprint arXiv:2303.17276*.

Lewis, D. K. (1969). *Convention: A philosophical study*. John Wiley & Sons.

Liu, R., & Shah, N. B. (2023). *ReviewerGPT? An exploratory study on using large language models for paper reviewing.*

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *ArXiv*, *abs/2304.03442*.

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., & Bernstein, M. S. (2022). Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th annual acm symposium on user interface software and technology* (pp. 1–18).

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Qin, Z., et al. (2023). *Large language models are effective text rankers with pairwise ranking prompting.*

Reiss, M. V. (2023, April). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark.* arXiv. Retrieved 2024-12-20, from http://arxiv.org/abs/2304.11085 (arXiv:2304.11085 [cs]) doi: 10.48550/arXiv.2304.11085

Salinas, A., & Morstatter, F. (2024, April). *The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance.* arXiv. Retrieved 2024-12-11, from http://arxiv.org/abs/2401.03729 (arXiv:2401.03729 [cs]) doi: 10.48550/arXiv.2401.03729

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature reviews neuroscience*, *8*(9), 657–661.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, *1*(2), 143–186.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Techical Journal*, *27*(3), 379–423.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Blackwell.

Stalnaker, R. C. (1978). Assertion. *Pragmatics*, 315-332.

Szpunar, K. K. (2010). Episodic future thought: An emerging concept. *Perspectives on Psychological Science*, *5*(2), 142–162.

Touvron, H., et al. (2023). *Llama 2: Open foundation and fine-tuned chat models.*

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., ... Sui, Z. (2023). *Large language models are not fair evaluators.*

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... Zhou, D. (2023). *Self-consistency improves chain of thought reasoning in language models.*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., ... others (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes*, *93*(1), 1–13.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.