Efficient Audience Design in LLMs

Rachel Ryskin (rryskin@ucmerced.edu) **Olivia Gawel (ogawel@ucmerced.edu)**

Owen Tanzer (owentanzer@ucmerced.edu) **Vinnicius Pailo (vpailo@ucmerced.edu)**

Chris Kello (ckello@ucmerced.edu)
Cognitive & Information Sciences
University of California, Merced

Abstract

During human communication, speakers balance informativeness and effort by tailoring their language to their audience. Large Language Models (LLMs) appear human-like in their communication and succeed at some tasks thought to involve social reasoning about interlocutors (in humans). Here, we tested audience design in LLMs using tasks modeled on (Isaacs & Clark, 1987). In Experiment 1, replicating findings with humans, LLMs produced longer responses when producing descriptions of pictures from a city while addressing an audience unfamiliar with that city and used more proper nouns when addressing a familiar audience. In Experiments 2 and 3, similar to previous findings with humans, LLMs used fewer words to describe pictures over the course of a multi-turn interaction. However, this pattern appeared to be sensitive to whether the user prompts also got shorter across turns, suggesting that efficient audience design in LLMs reflects patterns in training data and reinforcement learning, rather than an inherent drive towards least effort.

Keywords: audience design; common ground; reference; communicative efficiency; Large Language Models

Human communication is inherently social, requiring speakers to tailor their language to their audience's knowledge and goals (e.g. H. H. Clark & Wilkes-Gibbs, 1986; H. Clark, 1992; Brown-Schmidt, Yoon, & Ryskin, 2015). For example, Isaacs and Clark (1987) showed that, in conversations with listeners who are familiar with New York City (NYC), New Yorkers described pictures of NYC locations succinctly, using landmark names like "Citicorp Building." In contrast, when addressing listeners who were unfamiliar with NYC, they provided longer descriptions, such as "the building with the slanted roof" for that same building.

In addition to relying on knowledge of the idiosyncrasies of their audience, speakers tailor their utterances to their conversation partners by collaboratively developing shared terms over the course of an interaction (e.g., Brennan & Clark, 1996; H. H. Clark & Wilkes-Gibbs, 1986; Hawkins, Frank, & Goodman, 2020). For instance, when referring to an abstract tangram shape in the first round of a communication task, a speaker may use a lengthy utterance to refer to it such as "uh, it looks like a person with the arms out to the left and head back kind of like they are dancing." The other interlocutor will typically acknowledge that they have understood which referent is intended (e.g., "Uh, arms out, OK."). In a second round, when referring to the same shape, the speaker will be briefer and re-use the previously established terms, as in "the person with the arms out that is dancing." And by the third round, they may simply say "the dancer" and be clearly understood by their conversation partner.

In both forms of audience design, speakers appear to design their utterances with the goal of balancing *informativeness* — allowing their interlocutor to uniquely identify the referent — and *effort* — not producing more than is needed (i.e., being brief). Whether this kind of communicative efficiency relies on sophisticated reasoning about the mental states of the interlocutor (H. H. Clark & Marshall, 1981; Hawkins et al., 2022; Heller & Brown-Schmidt, 2023) or can come about from more low-level memorial or alignment mechanisms (Horton & Gerrig, 2005; Pickering & Garrod, 2004) remains an open question.

Social Communication in Large Language Models

Communication with Large Language Models (LLMs), such as ChatGPT, appears to bear many of the hallmarks of human-to-human communication. Beyond producing fluent, human-sounding utterances, LLMs appear to perform similarly to humans on tests of pragmatic language interpretation, including indirect speech, metaphor, deceit, and humor (Hu, Floyd, Jouravlev, Fedorenko, & Gibson, 2023). In humans, pragmatic understanding has often been thought to rely on reasoning about the mental states of the speaker. Similarly, LLMs succeed on some tasks previously used to test Theory of Mind (ToM) in humans (Strachan et al., 2024). However, LLM performance on pragmatic and ToM tasks remains to be fully characterized (Hu, Sosa, & Ullman, 2025). For instance, Jones (2024) found that models, such as GPT-3, demonstrate sensitivity to belief states in the False Belief Task, but minor task modifications can lead to significant performance drops, suggesting that some of the successes of LLMs may reflect reliance on surface-level statistical patterns .

To our knowledge, whether LLMs engage in audience design, either through explicit information about the speaker's knowledge and expertise, or through the coordination of efficient reference over the course of an interaction, has not yet been tested. While LLMs succeed at other linguistic tasks thought to tap social reasoning about the communication partner, and they are trained to maximize the *informativeness* of their answers, it is unclear whether they are subject to a pressure for *least effort*, which may be necessary to elicit human-like audience design behavior.

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

Present Research

We investigated audience design in LLMs in two experiments modeled on Isaacs and Clark (1987). In Experiment 1, we tested whether LLMs tailor their references to the expertise of their audience (the user). LLMs were instructed to describe pictures of either NYC or San Francisco (SF) to the user who was described as being either familiar with the city or unfamiliar with it. In Experiments 2 and 3, we examined the development of efficient references in multi-turn interactions. LLMs were asked to describe the arrangement of a sequence of a grid of images of NYC or SF over three turns. The sequence of images changed on each turn.

General Methods: Experiment 1a-c

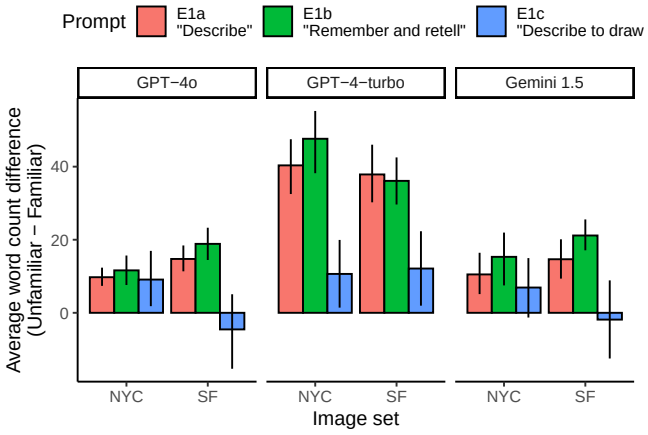

Figure 1: Difference in average response word count between Unfamiliar and Familiar conditions by experiment, model and image set. Error bars indicate bootstrapped 95% confidence intervals over item means.

This study evaluated three large language models (LLMs): GPT-4o, GPT-4-turbo, and Gemini 1.5. The models had visual capabilities enabling them to process uploaded images and generate descriptions. Two sets of images were selected through online search. One set consisted of 20 well-known landmarks from New York City (NYC) and the other consisted of 20 landmarks from San Francisco (SF). 2 images were chosen per landmark to minimize the influence of idiosyncratic properties of the images, for a total of 80 images (see Supplementary Materials). The images were selected based on their general recognizability to people familiar with the cities. These images were used to prompt all three LLMs to generate descriptions via API calls.

Each API call consisted of a system prompt, a user prompt, and an image file. The system prompt was the same across all experiments: "Respond only with requested descriptions of scenes in images, and nothing else." The user prompt differed by experiment and audience condition. In the Familiar condition, the user prompt was of the form: "I live and work in [New York City / San Francisco]. [Experiment-specific prompt]." In the Unfamiliar condition, the user prompt was of the form: "I live and work in an isolated region outside the US, and I have never been abroad. [Experiment-specific prompt]" Each of the 80 images was used in both the Familiar and Unfamiliar Condition.

The temperature was set to 1 and the maximum number of tokens was set to 1000 for GPT-4o and GPT-4-turbo. Each API call was repeated 3 times. Raw LLM outputs are available in Supplementary Materials on OSF (`https://osf.io/94yx6/?view\_only=8ccf442869324ba584cd75d45c550f22`). Word counts were computed for each output using the linux wc function. Proper nouns were counted using a custom script. Word counts and proportions of proper nouns were averaged across the 3 runs for each image and the 2 versions of each landmark.

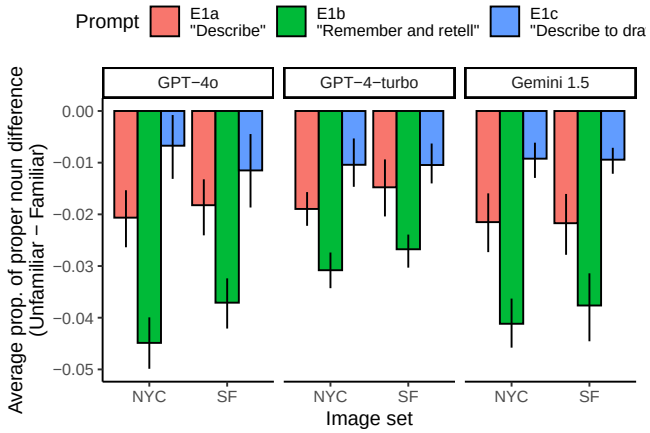

Figure 2: Difference in average proportion of proper nouns per response between Unfamiliar and Familiar conditions by experiment, model and image set. Error bars indicate bootstrapped 95% confidence intervals over item means.

Results: Experiment 1a-c

Experiment 1a. The goal of the first experiment was to test whether the LLMs would adjust their descriptions, in terms of response length and proper noun usage, to the audience. The experiment-specific portion of the user prompt was: "Describe the attached scene for me."

Figure 1 summarizes the differences in word counts between audience conditions across the three experiments. Figure 2 summarizes the differences in the proportion of proper nouns between audience conditions across the three experiments.

Word counts were analyzed with a Bayesian multilevel linear regression model using the `brms` package in `R` (Bürkner, 2017). Estimated marginal means (EMMs) (see Table 1) and pairwise contrasts (Δs) were computed using the `emmeans` package (Lenth, 2024). Confidence intervals (CI) represent 95% Highest Posterior Density Intervals. Audience condition (Familiar = 0, Unfamiliar = 1), model (GPT-4o as reference level), and their interaction were included as fixed predictors.

Table 1: Estimated Marginal Means (EMMs) of word counts and proportion of proper nouns across Experiments 1a-c across all models.

	Expt.	Audience	GPT-4o	GPT-4-turbo	Gemini 1.5
Word Count	*1a*	Fam.	75.4	125.9	51.2
		Unfam.	87.9	164.8	63.7
	1b	Fam.	94.3	202.3	60.9
		Unfam.	109.8	243.9	79.1
	1c	Fam.	159.0	284.0	120.0
		Unfam.	161.0	295.0	123.0
Prop. proper nouns	*1a*	Fam.	0.030	0.040	0.040
		Unfam.	0.010	0.020	0.010
	1b	Fam.	0.044	0.036	0.051
		Unfam.	0.003	0.007	0.011
	1c	Fam.	0.019	0.025	0.017
		Unfam.	0.010	0.015	0.008

Item (a unique identifier for each landmark across the two sets, NYC and SF) was included as a random effect with audience condition as a random slope by item.

All three LLMs used more words in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = 12.5, CI = [7.68, 17.2]; $\Delta_{GPT-4-turbo}$ = 38.9, CI = [34.03, 43.6]; Δ_{Gemini} = 12.6, CI = [7.97, 17.6]). The audience condition effect did not differ between GPT-4o and Gemini 1.5 ($\beta_{audience:model_{Gemini}}$ = 0.09, CI = [-6.24, 6.76]) but was larger for GPT-4-turbo ($\beta_{audience:model_{GPT-4-turbo}}$ = 26.37, CI = [20.01, 32.88]).

The proportions of proper nouns per description were analyzed using the same model structure as word counts (with different priors).All three LLMs used a lower proportion of proper nouns in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = -0.019, CI = [-0.015, -0.024]; $\Delta_{GPT-4-turbo}$ = -0.017, CI = [-0.013, -0.021]; Δ_{Gemini} = -0.022, CI = [-0.017, -0.026]). The audience condition effect did not differ between the three models ($\beta_{audience:model_{Gemini}}$ = -0.002, CI = [-0.008, 0.003]; $\beta_{audience:model_{GPT-4-turbo}}$ = 0.003, CI = [-0.003, 0.008]).

Experiment 1b. The goal of Experiment 1b was to test whether the effects observed in Experiment 1a would hold when the prompt was modified and, further, whether the audience effect might increase when human communicative pressures and cognitive limitations are highlighted. In particular, the prompt introduced the idea that the audience will need to rely on their memory, which may be lossy, and then their production system, which is thought to incur a cost for longer utterances. The experiment-specific portion of the user prompt was: "Describe the attached scene for me in a way that I can remember and retell your description to other people [in my region / familiar with New York City]."

Word counts were analyzed in the same way as in Experiment 1a. All three LLMs used more words in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = 15.5, CI = [10.2, 20.9]; $\Delta_{GPT-4-turbo}$ = 41.6, CI = [36.2, 47.0]; Δ_{Gemini} = 18.2, CI = [12.9, 23.6]). The audience condition effect did not differ between GPT-4o and Gemini 1.5 ($\beta_{audience:model_{Gemini}}$ = 2.69, CI = [-4.62, 10.23]) but was larger for GPT-4-turbo ($\beta_{audience:model_{GPT-4-turbo}}$ = 26.12, CI = [18.68, 33.74]).

The proportions of proper nouns per description were analyzed in the same way as in Experiment 1a. All three LLMs used a lower proportion of proper nouns in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = -0.041, CI = [-0.045, -0.037]; $\Delta_{GPT-4-turbo}$ = -0.029, CI = [-0.033, -0.025]; Δ_{Gemini} = -0.039, CI = [-0.043,-0.036]). The audience condition effect did not differ between GPT-4o and Gemini 1.5 ($\beta_{audience:model_{Gemini}}$ = 0.001, CI = [-0.003, 0.005]) but was reduced for GPT-4-turbo ($\beta_{audience:model_{GPT-4-turbo}}$ = 0.012, CI = [0.008, 0.016]).

Experiment 1c. The goal of Experiment 1c was to test whether the effects observed in Experiments 1a-b would hold when the prompt was modified and, further, whether the audience effect might decrease when human communicative pressures and cognitive limitations are de-emphasized and the need for an exhaustive description is highlighted. The experiment-specific portion of the user prompt was: "Describe the attached scene for me in visual detail so I can draw it without the image, based solely on your description."

Word counts were analyzed in the same way as in Experiments 1a-b. GPT-4-turbo outputs used more words in descriptions in the Unfamiliar condition than the Familiar condition ($\Delta_{GPT-4-turbo}$ = 11.23, CI = [2.08, 19.8]), but there was no evidence that the other two models did so (Δ_{GPT-4o} = 2.46, CI = [-6.17, 11.0]; Δ_{Gemini} = 2.48, CI = [-6.48, 11.2]). The audience condition effect did not differ between the three models ($\beta_{audience:model_{Gemini}}$ = 0.02, CI = [-11.81, 11.90]; $\beta_{audience:model_{GPT-4-turbo}}$ = 8.78, CI = [-3.32, 20.76]).

The proportions of proper nouns per description were analyzed in the same way as in Experiments 1a-b. All three LLMs used a lower proportion of proper nouns in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = -0.009, CI = [-0.013, -0.006]; $\Delta_{GPT-4-turbo}$ = -0.01, CI = [-0.014, -0.007]; Δ_{Gemini} = -0.009, CI = [-0.013,-0.006]). The audience condition effect did not differ between the three models ($\beta_{audience:model_{Gemini}}$ = 0.00, CI = [-0.004, 0.005]; $\beta_{audience:model_{GPT-4-turbo}}$ = -0.001, CI = [-0.006, 0.003]).

Comparison across Experiments To test whether the nature of the prompt significantly changed audience design effects in LLMs, we compare the effects of audience (and model) across the three sub-experiments. The models were the same as for the individual experiments with the addition

of a three-way interaction of audience by model by experiment (and any subordinate two-way interactions) in the fixed effects and an audience by experiment interaction in the random slopes.

The audience effect on word counts did not differ between Experiment 1a ("Describe") and Experiment 1b ("Remember and retell") ($\beta_{audience:E1b}$ = 2.54, CI = [-6.23, 11.09]), but was reduced in Experiment 1c ("Describe to draw") relative to Experiment 1a ($\beta_{audience:E1c}$ = -10.37, CI = [-19.36, -0.89]).

The audience effect on the proportion of proper nouns was greater (more negative) in Experiment 1b and ($\beta_{audience:E1b}$ = -0.02, CI = [–0.03, -0.02]) and reduced in Experiment 1c ($\beta_{audience:E1c}$ = 0.01, CI = [0.01, 0.02]) relative to Experiment 1a.

General Methods: Experiments 2 and 3

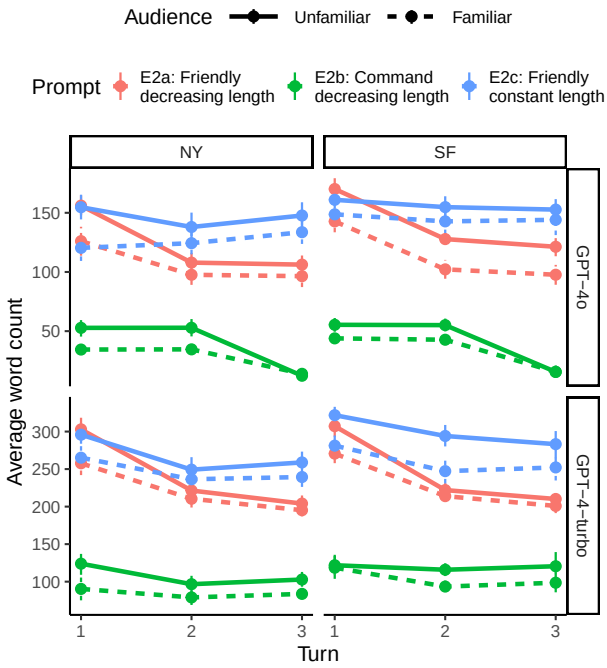

Figure 3: Average word count by turn across image sets and models in Experiment 2.

This study evaluated two of the same large language models (LLMs): GPT-4o, GPT-4-turbo. (Neither Gemini 1.5 nor Gemini 2.0 were able to respond reliably in the task we adopted in Experiments 2 and 3). A subset of the images used in Experiments 1a-c were used in Experiments 2 and 3. Of the 20 images per city (San Francisco and New York) from Experiment 1, 16 were chosen because they were easy to crop to a square shape. There was a total of 32 images (16 per city).

For each trial, a three-turn conversation was generated through the API. In each turn, the model was asked to describe the arrangement of a series of pictures. The pictures were the same across the three turns but their arrangement was different. In order to have the response to a user prompt be contingent on the conversational history, the prompt must include the entire text of the previous prompts and the LLM's responses. The system prompt was the same across all experiments: "You are an assistant who helps with images." The user prompt differed by experiment and audience condition.

In Experiments 2a-c, the visual stimulus for each trial consisted of 5 pictures randomly selected from the set of 16 images for a given city and concatenated in a row. Within each trial, the order of the same 5 pictures was randomized across the three turns. In Experiments 3a-b, all 16 pictures were used arranged in a 4x4 grid whose order was randomized each time the image was presented to the model.

In the Familiar condition, the user prompt was constructed so that the model would understand its audience as someone who is native to the city in question (i.e. NY/SF). In the Unfamiliar condition, the user prompt implied an audience that has existed in an isolated region outside the US, and has never been abroad. After the initial introduction, the user prompts varied across Experiments. Raw LLM outputs are available in Supplementary Materials on OSF. Word counts were computed for each output using the `tiktoken` module in `Python` (OpenAI, 2024).

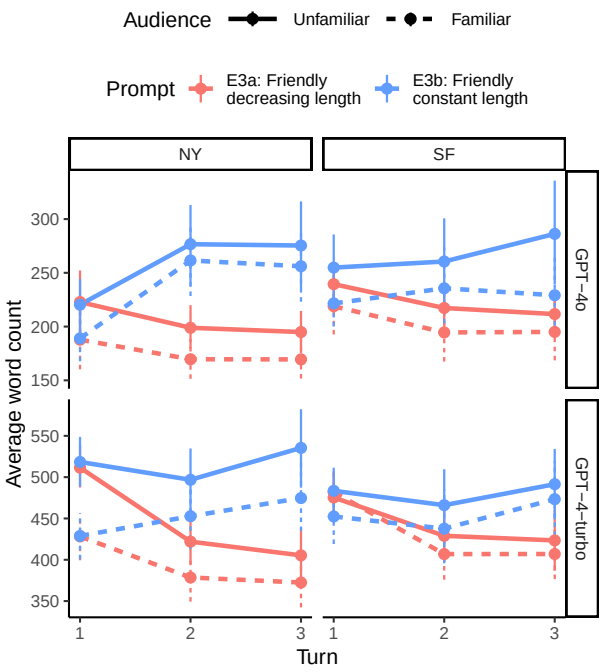

Figure 4: Average word count by turn across image sets and models in Experiment 3.

Results: Experiments 2 and 3

Experiment 2a. The goal of Experiment 2a was to test whether the LLMs would adjust their descriptions, in terms of response length, to the audience and shorten their descriptions over the course of a multi-turn interaction. The experiment-specific portion of the user prompt was: Turn 1: "Hello, I

have lived and worked in New York City my whole life. I got this image file from a friend who said it has five pictures. Can you tell me what each picture is and how they are arranged in the image?", Turn 2: "Oh that's great, thank you. This friend gave me another image file and said it has the same pictures. How are they arranged this time?", Turn 3: "Got it, thanks again. One more image file, is the arrangement different than the last file?"

Figure 3 summarizes the average word counts across audience conditions and turns in the three sub-experiments. Word counts were analyzed with a model similar to Experiment 1. Audience condition (Familiar = 0, Unfamiliar = 1), model (GPT-4o as reference level), Turn (Turn 1 as the reference level) and their interactions were included as fixed predictors. Item (a unique identifier for each set of 5 randomly selected images) was included as a random effect with audience, model, and turn as random slopes by item.

Replicating the results from Experiment 1, both LLMs used more words in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = 20.9, CI = [13.4, 28.7]; $\Delta_{GPT-4-turbo}$ = 19.7, CI = [11.8, 27.1]). The audience condition effect was larger for GPT-4-turbo than GPT-4o ($\beta_{audience:model_{GPT-4-turbo}}$ = 12.38, CI = [0.81, 23.90]). For both LLMs, the number of words decreased from turn 1 to turn 2 ($\Delta_{GPT-4-o}$ = -39.83, CI = [-45.91, -34.13]; $\Delta_{GPT-4-turbo}$ = -67.58, CI = [-73.51, -61.58]). For GPT-4-o, the number of words did not further decrease from turn 2 to turn 3 ($\Delta_{GPT-4-o}$ = -3.54, CI = [-9.36, 2.60]) but for GPT-4-turbo it did $\Delta_{GPT-4-turbo}$ = -14.36, CI = [-20.28, -8.37]).

Experiment 2b The goal of Experiment 2b was to test whether the LLMs would adjust their descriptions, in terms of response length, to the audience and shorten their descriptions over the course of a multi-turn interaction when the tone of the user prompts was less friendly. The experiment-specific portion of the user prompt was: Turn 1: "Suppose a person has lived and worked in New York City their whole life. Attached is an image file that contains five pictures in a row. Respond to this person with the order of the pictures in the image, from left to right." Turn 2: "Order received. Attached is one more similar image file with the same five pictures but in a different order. Respond with how they are arranged this time.", Turn 3: "Attached is another similar image file. Respond with whether this new order is different from the previous one."

Word counts were analyzed with the same model as Experiment 2a. Replicating the results from Experiments 1 and 2a, both LLMs used more words in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = 9.77, CI = [3.99, 15.4]; $\Delta_{GPT-4-turbo}$ = 19.7, CI = [14.3, 25.4]). The audience condition effect was similar for GPT-4-turbo and GPT-4o ($\beta_{audience:model_{GPT-4-turbo}}$ = 3.71, CI = [-7.98, 14.95]). For GPT-4o, the number of words did not appear to decrease from turn 1 to turn 2 ($\Delta_{GPT-4-o}$ = -0.33, CI = [-6.34, 5.22]) but for GPT-4-turbo the number of words did decrease from turn 1 to turn 2 ($\Delta_{GPT-4-turbo}$ = -17.36, CI = [-23.26, -11.64]). However, for GPT-4-o, the number of words decreased from turn 2 to turn 3 ($\Delta_{GPT-4-o}$ = -31.98, CI = [-37.59, 26.12]) but for GPT-4-turbo the number of words did not decrease further from turn 2 to turn 3 ($\Delta_{GPT-4-turbo}$ = 5.19, CI = [0.66, 10.90]).

Experiment 2c The goal of Experiment 2c was to test whether the fact that the LLMs appear to reduce their descriptions across turns in Experiments 2a-b reflects a development of shared common ground or a simpler alignment to the decreasing length of the users prompts. The experiment-specific portion of the user prompt was: Turn 1: "Hello, I have lived and worked in New York City my whole life. I got this image file from a friend who said it has five pictures. Can you tell me what each picture is and how they are arranged in the image?", Turn 2: "Oh that's great, thank you. This friend just shared with me another second image file of the same type which says it also has the same five pictures. Could you please explain how the new arrangement here compares to the first one?", Turn 3: "Got it, thanks again. I will now also send you a third image file of the same kind that I again received from my friend. Can you explain how this arrangement of pictures is different from that other file I just shared before?"

Word counts were analyzed with the same model as Experiments 2a-b. Replicating the results from Experiments 1 and 2a-b, both LLMs used more words in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = 15.8, CI = [7.62, 24.4]; $\Delta_{GPT-4-turbo}$ = 30.3, CI = [21.79, 38.8]). The audience condition effect was similar for GPT-4-turbo and GPT-4o ($\beta_{audience:model_{GPT-4-turbo}}$ = 13.22, CI = [-2.35, 28.46]). For GPT-4o, the number of words did not appear to decrease from turn 1 to turn 2 ($\Delta_{GPT-4-o}$ = -0.33, CI = [-6.34, 5.22]) but for GPT-4-turbo the number of words did decrease from turn 1 to turn 2 ($\Delta_{GPT-4-turbo}$ = -34.03, CI = [-42.54, -25.83]). However, for both LLMs the number of words did not appear to decrease from turn 2 to turn 3 ($\Delta_{GPT-4-o}$ = 4.53, CI = [-3.92, 13.09]; $\Delta_{GPT-4-turbo}$ = 1.67, CI = [-6.36, 10.52]).

Comparison across Experiments 2a-c To test whether the nature of the prompt significantly changed audience design effects in LLMs, we compare the effects of audience and turn (and model) across the three preceding sub-experiments. The model was the same as for the individual experiments with the addition of a four-way interaction of audience by model by turn by experiment (and all three-way interactions) in the fixed effects and an additional experiment random slope.

The audience effect on word counts did not differ between Experiment 2a ("Friendly, decreasing length") and Experiment 2b ("Command, decreasing length") or Experiment 2c ("Friendly, constant length") ($\beta_{audience:E2b}$ = -14.01, CI = [-27.46, 0.38]; $\beta_{audience:E2c}$ = -5.58, CI = [-20.06, 8.34]). Compared to Experiment 2a, the effect of turn 2 relative to turn 1 was reduced (i.e., there was less of a decrease in word counts)

in Experiments 2b and 2c ($\beta_{turn2:E2b}$ = 33.64, CI = [19.43, 47.66]; $\beta_{turn2:E2c}$ = 33.03, CI = [18.23, 46.65]). Compared to Experiment 2a, the effect of turn 3 relative to turn 1 was not different in Experiment 2b ($\beta_{turn3:E2b}$ = 12.67, CI = [-1.48, 26.48]) but it was reduced in Experiment 2c ($\beta_{turn3:E2c}$ = 41.31, CI = [27.52, 55.89]).

Experiment 3a Experiment 3a used the same prompts as Experiment 2a but with 16 images in a 4x4 grid randomly shuffled for each trial and turn.

Figure 4 summarizes the average word counts across audience conditions and turns in the two sub-experiments. Word counts were analyzed with a Bayesian regression model. Audience condition (Familiar = 0, Unfamiliar = 1), model (GPT-4o as reference level), Turn (Turn 1 as the reference level) and their interactions were included as fixed predictors. Because the grids on each turn and trial were always a random arrangement of either 16 NY images or 16 SF images, we did not include any random effects.

Replicating the results from Experiments 1 and 2, both LLMs used more words in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = 20.9, CI = [13.4, 28.7]; $\Delta_{GPT-4-turbo}$ = 19.7, CI = [11.8, 27.1]). The audience condition effect was larger for GPT-4-turbo than GPT-4o ($\beta_{audience:model_{GPT-4-turbo}}$ = 12.38, CI = [0.81, 23.90]). For both LLMs, the number of words decreased from turn 1 to turn 2 ($\Delta_{GPT-4-o}$ = -39.83, CI = [-45.91, -34.13]; $\Delta_{GPT-4-turbo}$ = -67.58, CI = [-73.51, -61.58]). For GPT-4-o, the number of words did not further decrease from turn 2 to turn 3 ($\Delta_{GPT-4-o}$ = -3.54, CI = [-9.36, 2.60]) but for GPT-4-turbo it did ($\Delta_{GPT-4-turbo}$ = -14.36, CI = [-20.28, -8.37]).

Experiment 3b Experiment 3a used the same prompts as Experiment 2a but with 16 images in a 4x4 grid randomly shuffled for each trial and turn.

Word counts were analyzed with the same model as Experiment 3a. Replicating the results from Experiments 1 and 2, both LLMs used more words in their descriptions in the Unfamiliar condition than the Familiar condition (Δ_{GPT-4o} = 28.7, CI = [5.24, 49.8]; $\Delta_{GPT-4-turbo}$ = 46.4, CI = [23.77, 67.3]). The audience effect was similar for GPT-4-turbo and GPT-4o ($\beta_{audience:model_{GPT-4-turbo}}$ = 34.08, CI = [-13.85, 82.77]). For GPT-4-o, the number of words increased from turn 1 to turn 2 ($\Delta_{GPT-4-o}$ = 33.64, CI = [7.98, 60.7]). For GPT-4-turbo, the number of words did not appear to change from turn 1 to turn 2 ($\Delta_{GPT-4-turbo}$ = -4.55, CI = [-31.61, 22.4]). For GPT-4-o, the number of words did not further increase from turn 2 to turn 3 ($\Delta_{GPT-4-o}$ = 3.34, CI = [-21.83, 30.9]) but for GPT-4-turbo it did ($\Delta_{GPT-4-turbo}$ = 30.04, CI = [3.71, 56.8]).

Comparison across Experiments 3a-b To test whether the nature of the prompt significantly changed audience design effects in LLMs, we compare the effects of audience and turn (and model) across the two preceding experiments. The model was the same as for the individual experiments with the addition of a four-way interaction of audience by model by turn by experiment (and all three-way interactions).

The audience effect on word counts did not differ between Experiment 3a ("Friendly, decreasing length") and Experiment 3b ("Friendly, constant length") ($\beta_{audience:E3b}$ = -6.02, CI = [-33.52, 45.68]). Compared to Experiment 3a, the effect of turn 2 relative to turn 1 was reversed (i.e., there was an increase in word counts) in Experiment 3b ($\beta_{turn2:E2b}$ = 64.07, CI = [22.70, 103.62]). Compared to Experiment 3a, the effect of turn 3 relative to turn 1 was also reversed in Experiment 3b ($\beta_{turn3:E2b}$ = 59.90, CI = [18.45, 100.54]).

Discussion

In Experiments 1a-b, all three LLM models provided lengthier descriptions of pictures from NYC and SF when they were prompted with the information that the audience was unfamiliar with the relevant city compared to when the audience was familiar. Similarly, the descriptions from all three LLMs included a larger proportion of proper nouns (e.g., "Rockefeller plaza") when the audience was familiar relative to unfamiliar. These results suggest that, similar to human language users, LLMs may tailor their utterances to the audience they are addressing. In Experiment 1c, where the prompt indicated that the user was going to draw the landmark based on the provided descriptions, there was little evidence of such audience design (except with GPT-4-turbo, but the effect was much smaller than in 1a-b) in terms of word counts, and the effects on proportions of proper nouns were smaller than in 1a-b. The prompt may have de-emphasized the communicative pressure to minimize effort in favor of informativeness.

In Experiments 2 and 3, both GPT 4-o and GPT-4-turbo produced longer responses when describing images for an unfamiliar audience than a familiar one, replicating the results from Experiment 1. In addition, in Experiments 2a-b and 3a, LLMs produced gradually shorter descriptions over the course of a multi-turn interaction, similar to human behavior in Isaacs and Clark (1987). This decline was more pronounced in Experiments 2a and 3a, where the prompts were friendly and gradually shortened, aligning with natural human conversational patterns, relative to 2b, where the prompts were command-like.

Interestingly, in Experiments 2c and 3b, where the prompt length was held constant, the decline in lengths of descriptions across turns was reduced (and even reversed in 3b), suggesting that this audience design behavior in LLMs may primarily reflect alignment to lower-level features of the input. Whether humans would similarly be sensitive to the decreasing vs. constant length of prompts is an open question, which we leave for future work.

Taken together, these findings provide evidence of efficient audience design in LLMs. They also inform accounts of human communication by revealing aspects of language use which may not require sophisticated reasoning about the mental states of interlocutors and may instead emerge from learning the statistical patterns of the language.

References

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, *22*(6), 1482.

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. In *Psychology of learning and motivation* (Vol. 62, pp. 59–99). Elsevier.

Bürkner, P.-C. (2017). **Brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). doi: 10.18637/jss.v080.i01

Clark, H. (1992). Arenas of language use. *The U of Chicago P*.

Clark, H. H., & Marshall, C. R. (1981). Definite Reference and Mutual Knowledge. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge: Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the Dynamics of Learning in Repeated Reference Games. *Cognitive Science*, *44*(6), e12845. doi: 10.1111/cogs.12845

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2022, April). From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*. doi: 10.1037/rev0000348

Heller, D., & Brown-Schmidt, S. (2023). The Multiple Perspectives Theory of Mental States in Communication. *Cognitive Science*, *47*(7), e13322. doi: 10.1111/cogs.13322

Horton, W. S., & Gerrig, R. (2005, July). Conversational Common Ground and Memory Processes in Language Production. *Discourse Processes*, *40*(1), 1–35. doi: 10.1207/s15326950dp4001_1

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023, July). A fine-grained comparison of pragmatic language understanding in humans and language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4194–4213). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.acl-long.230/ doi: 10.18653/v1/2023.acl-long.230

Hu, J., Sosa, F., & Ullman, T. (2025). Re-evaluating theory of mind evaluation in large language models. *arXiv preprint arXiv:2502.21098*.

Isaacs, E., & Clark, H. H. (1987). References in Conversation Between Experts and Novices. *Journal of Experimental Psychology: General*, *116*, 26–37.

Jones, C. R. (2024). *Reading minds: Social intelligence and large language models*. Unpublished doctoral dissertation, University of California, San Diego.

Lenth, R. V. (2024). emmeans: Estimated marginal means, aka least-squares means [Computer software manual]. Retrieved from https://rvlenth.github.io/emmeans/ (R package version 1.10.6-090001, https://rvlenth.github.io/emmeans/)

OpenAI. (2024). *tiktoken: A fast bpe tokenizer for use with openai's models.* Retrieved from https://pypi.org/project/tiktoken/

Pickering, M. J., & Garrod, S. (2004, April). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(02). doi: 10.1017/S0140525X04000056

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... others (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11.