ExpressivityBench: Can LLMs Communicate Implicitly?

Joshua Tint[*♡], Som Sagar[♠], Aditya Taparia[♠], Kelly Raines[♠],
Bimsara Pathiraja[♠], Caleb Liu[♠], and Ransalu Senanayake[♠]

[♡]Amazon
[♠]Arizona State University
{jrtint, ssagar6, ataparia, kraine5, bpathir1, calebliu, ransalu}@asu.edu

Abstract

Human communication is often implicit, conveying tone, identity, and intent beyond literal meanings. While large language models have achieved strong performance on explicit tasks such as summarization and reasoning, their capacity for expressivity, or implicit communication, remains underexplored. We introduce **ExpressivityBench**, a framework for evaluating the expressivity of LLMs using information-theoretic communication models. Our approach quantifies how well LLM-generated text communicates target properties without explicit mention, across nine tasks spanning emotion, identity, and tone. To enable scalable and reproducible evaluation, we employ LLM-based graders validated against human judgments. Our results reveal that while models are adept at expressing affective content, they struggle with sociolinguistic signals, lagging behind human baselines. This study provides a necessary step to evaluate human-like implicit communication, with implications for applications such as education, mental health support, and socially-aware dialogue systems. We provide code and data for our benchmark alongside our paper.

1 Introduction

Much of human communication is implicit. The phrasing and tone of a message can convey a number of topics beyond its literal meaning: shaping tone, signaling social identity, or subtly guiding interpretation (Knepper et al., 2017). When a doctor tells a patient, "You should make some lifestyle changes," versus, "There are some adjustments that could really benefit your health," the explicit message remains the same, but the implied tone and emotional impact differ. This ability to communicate implicitly is a core aspect of natural language use, yet it remains underexplored in Large Language Models (LLMs).

LLMs (OpenAI, 2023; Touvron et al., 2023) have transformed fields reliant on human communication, including education (OpenAI, 2023), customer support (Radford et al., 2019), legal services (Chern et al., 2024), and healthcare (Bubeck et al., 2023). Chatbots in these fields interact with humans in potentially sensitive and stressful situations, demanding user trust, which is most freely given if the LLM behaves in a human-like way (Huang et al., 2024; Ding et al., 2025). As models grow in size and capability, their performance is typically evaluated on explicit tasks like translation, summarization, and question-answering (Devlin et al., 2018; Brown et al., 2020), or on basic linguistic competence or reasoning tasks (Davies et al., 2023; Ziyu et al., 2023; Hao et al., 2024). However, true human-like communication requires more than factual accuracy—it requires *expressivity*, the ability to convey implicit information (Apresyan, 2018).

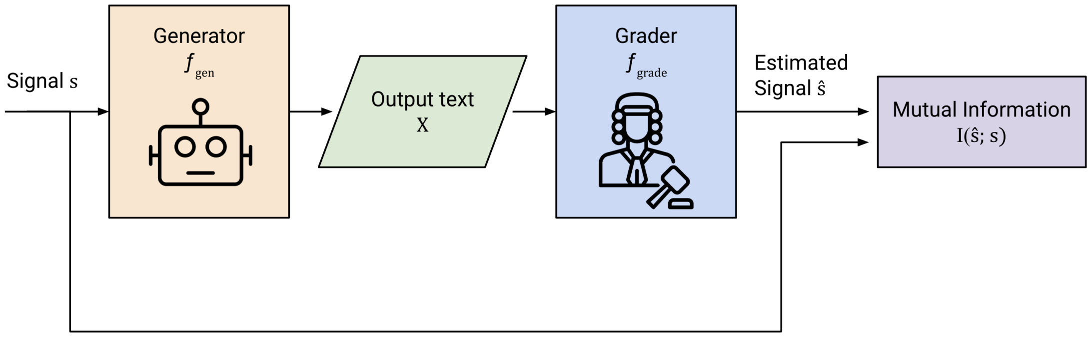
Expressivity influences not only tone but also social meaning. A model's word choice (e.g. formal, colloquial, or slang) can implicitly communicate a speaker's regional background, education level, or identity (Green, 2016). Recognizing this, developers have begun incorporating controls for persona-based expressivity; OpenAI, for example, introduced a feature into its ChatGPT models in Janurary 2025 allowing users to adjust "personalities" with settings like "chatty" or "Gen-Z" (Ferguson, 2025). Despite these advances, expressivity remains poorly understood in LLMs, in particular whether models can accurately convey these implicit tones and signals. Studying how models communicate implicitly is critical for improving their ability to generate "human-like" responses—enhancing trust, usability, and alignment with human expectations (Huang et al., 2024).

This study aims to quantitatively measure the ex-

[*]This work was completed prior to affiliation with Amazon.

Findings of the Association for Computational Linguistics: EACL 2026, pages 4500–4515

Figure 1: ExpressivityBench tests LLMs on their ability to implicitly express information using an information theoretic channel method, measuring a generator's ability to faithfully convey implicit signal to a grader.



pressivity of state-of-the-art LLMs. To this end, we focus on the following research question: **How expressive are LLMs when compared to humans?**

In order to answer this question, we present *ExpressivityBench*, a framework to evaluate expressivity in LLMs. We use information theoretic communication models to measure quantitatively the level of expressivity in LLM generated text. Our method employs a *grader* to evaluate (Benedetto et al., 2013) outputs generated by target LLMs on nine tasks. For reproducibility and scalability (Lee et al., 2023; Bai et al., 2024; Chern et al., 2024), we use LLMs as evaluators to determine how strongly implicit signals are expressed. We perform our task and grader selection through a human study and establish human-written text as a baseline, in order to understand how LLMs compare to humans. We find that LLMs perform strongly on commonplace expressivity tasks like emotional text generation, but perform much poorer on more complex identity or sociolinguistic-related tasks.

2 Related Work

2.1 Defining Expressivity

In this section, we explore the literature for definitions of expressivity related to natural language processing in order to formalize our own definition of expressivity. Study of expressivity in NLP is typically limited to emotions, as studied in affective computing devices (Pyreddy and Zaman, 2025; Hu et al., 2024). This includes recognizing and emulating emotions from written language (Plaza-del Arco et al., 2024). However, this limited focus does not capture other aspects that we use in our day-to-day communication, such as tone, metaphor, stylistic variation, and implicit social cues. We build a definition inclusive of this variation.

Thus, we adapt a definition from linguistics, to term "expressivity" as the state of communicating information implicitly: *showing, not telling* (Sanders; and Taniguchi, 2022). To further clarify, Yus distinguishes implicit from explicit communication by noting that implicit meaning must be inferred using context and pragmatics, whereas explicit meaning is directly represented in semantics (Yus, 1999). For example, "cheap" and "affordable" share a literal meaning but differ in connotation, while "greetings" conveys more formality than "hello." Since LLMs may struggle with contextual interpretation, studying expressivity helps reveal their linguistic limitations (Zhu et al., 2024). Other factors are often implicitly inferred as they are read, including sociolinguistic traits (Stecker, 2001). A writer's intended audience is often implicitly communicated as well, for instance one would speak differently to children rather than to adults (Gutt, 1996). From the literature, we note four broad categories of expressivity: communicating the speaker's intention (e.g. intended genre), speaker's audience (e.g. level of formality), speaker's identity (e.g. age and gender) and speaker's current state (e.g. current emotion) (Deckert and Kosecki, 2023; Choi et al., 2020).

2.2 Evaluating Large Language Models

Existing benchmarks for LLMs evaluate their performance across diverse tasks, including mathematics (Collins et al., 2023), logical reasoning (Parmar et al., 2024), and education (Dai et al., 2023). These benchmarks typically rely on either automated evaluation using external LLMs (Lin and Chen, 2023; Verga et al., 2024) or human feedback, as seen in Chatbot Arena (Chiang et al., 2024). Automated evaluation has gained traction due to its scalability and efficiency (Chang et al., 2024), and AI feedback has been proposed as a scalable alternative to RLHF (Tunstall et al., 2023; Lee et al., 2023). While automated evaluation can induce biases, it also improves reproducibility (Sharma et al., 2024). In order to mitigate weaknesses of automated evaluation, Laskar et al. (2024) recommends reporting detailed model versions, validating grader models on gold labels, and sharing evaluation data, which we practice for this study.

Evaluating language models' understanding of pragmatic communication and non-literal language remains challenging. Studies on figurative language comprehension have explored scene modeling for interpreting metaphors and idioms, and related tasks like dereferencing metonomies (Gu et al., 2022; Chakrabarty et al., 2022; Sravanthi et al., 2024). Benchmarks such as SocKET assess social communication aspects like humor and sarcasm (Choi et al., 2023), while DailyDialog evaluates conversational abilities (Li et al., 2017). Research has highlighted difficulties in teaching models empathy due to data limitations (Rashkin et al., 2019) and their tendency to favor literal interpretations over pragmatic understanding (Hu et al., 2023). However, most pragmatics-oriented benchmarks are largely discriminative, focusing on LLMs' ability to discern pragmatic information (Ma et al., 2025).

2.3 Constrained Generation Tasks

Constrained generation tasks assess models' ability to produce text which satisfies natural-language conditions, for example generating text which must implicitly communicate a signal. Recent work has introduced modular frameworks like COLLIE (Yao et al., 2023) to define compositional constraints across multiple linguistic levels, enabling more interpretable and flexible experimentation. Nevertheless, integrating complex constraints into LLMs remains a substantial challenge. Several studies highlight limitations in the ability of autoregressive models to maintain multiple constraints over longer generations (Chen and Wan, 2024; Garbacea and Mei, 2025), and the difficulty of balancing fluency and constraint satisfaction (Liu et al., 2024). These challenges are especially salient in expressive generation tasks, where models must not only satisfy abstract stylistic or sociolinguistic constraints but do so in a manner that appears natural and contextually appropriate to human readers.

3 ExpressivityBench

3.1 Overview

We take an information-theoretic approach to understanding expressivity, adopting a method derived from the "encoder-channel-decoder" model. Classically, these setups involve an "encoder" mechanism that communicates a signal through a channel, which must then be received and interpreted by a "decoder." The channel's capacity to communicate information can be derived from the similarity of the final transcribed message to its original signal. We adapt this to measure how well a model can communicate an implicit signal in a piece of natural language text. Formally, the goal of an expressivity experiment is to measure whether a *generator*, $f_{gen}(\cdot)$ can produce a piece of text in the domain d, implicitly containing a signal s strongly enough that it can be discerned by a *grader* $f_{grade}(\cdot)$ from a set of other signals $\mathcal{S}$.

To start, our benchmark has a domain d, which is a string representing the type of text that the model must generate (e.g. "song," or "recipe") and a signal category $\mathcal{S}$, a set which contains all possible signals which might be communicated in the experiment. A particular signal $s \in \mathcal{S}$ is selected, and the generator $f_{gen}(\cdot)$ is used to generate text with the following prompt W:

> "Write a $\langle d \rangle$ which conveys $\langle s \rangle$. Do not explicitly mention $\langle s \rangle$ in your response. Do not convey any of the following signals: $\langle \mathcal{S} \setminus s \rangle$"

The result, $X = f_{gen}(W)$, is the text that must be evaluated by the grader. To avoid unintentionally leaking s in the response, if X contains the signal s, then the response is regenerated. The blind grader, $f_{grade}(\cdot)$ is then prompted to select, of all the signals in $\mathcal{S}$, which one is present in X, using the following prompt:

> "Which of the following is conveyed in the text: $\langle \mathcal{S} \rangle$? Here is the text: $\langle X \rangle$ Now respond ONLY with one of the terms from this list (do not say none) and answer with no preamble: $\langle \mathcal{S} \rangle$?"

Its selection is the estimated signal $\hat{s}$. We repeat these experiments for a fixed number of times over all signals in $\mathcal{S}$ in order to create an induced probability distribution $p(s|\hat{s})$, i.e., the probability that, given the grader guessed some estimated signal $\hat{s}$, the intended signal was actually s. In order to measure how expressive the generator $f_{gen}(\cdot)$ is, we take the mutual information over this distribution $I(s;\hat{s})$ (A worked example of mutual information is in Appendix 9.4). This represents the number of bits of information that the signal s communicates towards the grader's guess, or equivalently, the number of bits of information that the grader's guess tells us about which signal was originally communicated. A high mutual information score indicates that the generator can communicate implicit signals with strong fidelity. We also compute the normalized mutual information, $N = I(s;\hat{s})/H(s)$, where $H(s)$ is the entropy of the ground truth labels. This normalized value represents the proportion of the total possible information that is actually attributable to the signal. A score of 1 would indicate perfect communication fidelity, while a score of 0 would indicate no signal transfer.

3.2 Human Study

In order to extend this experiment structure into a benchmark, we needed to select a set of tasks and a grader to evaluate with. In the interest of modeling human expressivity as closely as possible, we rooted these decisions in a human study that evaluated the ability of humans to recognize expressivity in a variety of tasks. This serves two purposes: understanding which types of expressive signals can be discerned by real humans, and understanding how closely various automated graders match real human graders' decisions.

To select an initial list of tasks for this human study, we sought out datasets of text annotated with signals. The criteria for inclusion were:

1. Text in the dataset had to be written by humans
2. The signal annotation had to be performed by humans
3. The type of signal used had to correspond to some dimension of expressivity: i.e., convey implicit information about the speakers' intention, audience, background or state. Many datasets fell under multiple of these categories.

We selected 10 datasets covering a diverse range of text domains and types of expressivity. Additionally, we were strongly motivated to include fiction genre as an expressivity category due to its prevalence in creative generation (Li et al., 2025), but we were unable to find a dataset that matched our criteria, so we collected one from the Goodreads website. Details about this dataset collection can be found in Appendix 9.2.

In the interest of keeping our tasks tractable and ensuring high automated grader quality, we limited the number of signals in each task. For tasks that had large numbers of signals annotated in the dataset (hobby, profession, genre, emotions, and tones), we used an algorithm described in Appendix 1 to extract a set of maximally distinct labels. For instance, the original Tone Analysis dataset had over 20 labels; we extracted 4. Some datasets had well over 5,000 entries (register, tones, skill level, political slant, age, and gender), so these had random samples of 500 texts taken.

For *emotion*, we drew from the PERC dataset, which contains poems labeled with one of seven emotions. There are many datasets of emotion-tagged text, but we chose this poetry-based one because poetry is typically thought of as a very expressive domain, and because it would add more diversity to the domains used in the benchmark (Sreeja and Mahalakshmi, 2019). The *register* task was constructed using a 500-sample subset of text messages from the Pavlick Formality Scores dataset, which contains messages annotated as "formal" or "casual" (Pavlick and Tetreault, 2016). Formality is particularly useful for analyzing stylistic nuance (Yus, 1999). For *genre*, we identified five literary categories from Goodreads, leveraging a random sample of 300 short story excerpts to assess narrative inference (see Appendix 9.2). To examine expressivity with respect to intended audience, we included an *MPAA* category task using 500 excerpts from the CLEAR Corpus, which consists of short stories and narrative texts written for U.S. schoolchildren of various ages (Crossley et al., 2021). Each excerpt is labeled with a rating analogous to the Motion Picture Association of America (MPAA) system: G (appropriate general audiences), PG (parental guidance suggested), or PG-13 (some material may be inappropriate for children under 13). For *tone*, we utilized 554 lines of dialogue labeled with pragmatic attitudes like cautionary, witty, or apologetic, sourced from the Tone Analysis dataset (Atif, 2023). In contrast to

Table 1: List of tasks used in the human study

Signal category	Signals	Text domain	# Samples	Source
emotion	sad, love, peace, joy, courage, surprise, hate	poem	450	[56]
register	formal, casual	text message	500	[46]
genre	science fiction, thriller, romance, humor, fantasy	short story excerpt	300	Goodreads
MPAA rating	PG, G, PG-13	short story excerpt	500	[16]
tone	cautionary, witty, apathetic, apologetic	line of dialogue	554	[3]
skill level	high writing quality, low writing quality	argumentative essay	500	[17]
political slant	left, center, right	news headline	500	[30]
age	author under 20 years old, author between 20 and 30 years old, author above 30 years old	blog post	500	[53]
gender	author is male, author is female	blog post	500	[53]
hobby (not used in benchmark)	author likes to cook, author likes to volunteer, author likes to hike, author likes to travel	line of dialogue	505	[64]
profession (not used in benchmark)	author is an engineer, author is a manager, author is a lawyer, author is a musician	line of dialogue	504	[64]

emotions, tone is a more intentional reflection of an author's attitude towards a subject (Greene, 2023). The *skill level* task was framed around writing quality, with samples from argumentative essays in the ASAP-2.0 dataset which were labeled according to grades received in an essay competition (Crossley et al.). In the original dataset, essays were scored from 1-6. We assigned natural-language labels "low quality writing" to essays with a score ≤ 2 and "high quality writing" to essays with a score ≥ 4. Essays with a score of 3 exactly were discarded. This categorization scheme corresponds as closely as possible with the top third and bottom third of essays. *Political slant* was assessed on 500 news headlines from the QBias dataset, which were labeled as leaning left, center, or right (Haak and Schaer, 2023). Our final four categories focus on testing the ability of models to infer personal information embedded in narrative voice. For *age* and *gender*, we used 500 blog posts each from the Blog Authorship Corpus, labeled according to three age brackets (under 20, 20–30, over 30) and binary gender (male, female), respectively (Schler et al., 2006). While many text datasets are annotated by gender and age, we chose this blogging-based corpus because blogging is a more personal medium, and therefore personal traits may be better reflected by it. Finally, we used the PersonaChat dataset for two classification tasks; this dataset consists of human-human conversations tagged by details of each interlocutor's persona. From these we created a *hobby* task, based on speakers' preferences like cooking or traveling, and *profession* such as engineer or musician (Zhang et al., 2018). Across all categories, we filtered out samples which contained the name of the signal in the text, matching the regeneration criteria for automated generation. A summary of the tasks can be found in Table 1. We provide all the human text samples as part of our benchmark along with our paper.

For our human study, each participant was presented with two random text samples from each of the 11 datasets. The questions in the study were resampled from a question pool of 50 per task so that each question would have the opportunity to be graded multiple times by multiple graders. The graders were asked to select, from each of the possible signals for that task, which one was conveyed in the sample. We presented our study to students at Arizona State University. All participants were fluent English speakers. 70% were in STEM programs, 8% were in humanities programs, and 22% were in business or finance programs. 62% were undergraduates, 16% were in Master's programs, and 22% were in PhD programs. We did not collect any other demographic information on participants. We collected 814 graded questions, 74 grades for each task. Among questions that had responses from three or more graders, our Fleiss's Kappa score was 0.85, indicating very strong agreement.

3.3 Task Selection

To identify tasks where expressivity could actually be perceived, we computed the normalized mutual information scores $N = I(\hat{s}; s)/H(s)$ between the grader's guesses and the ground truth signals. We rejected tasks where N was below 0.1. This eliminated two tasks: profession and hobby. Our complete results for this task can be seen in Table 2. Of the tasks that ultimately made it through this selection process, none had an N score below 0.15.

Table 2: Entropy scores for human graders across all tasks

Task	$I(s; \hat{s})$ (b)	$H(s)$ (b)	N
tone	1.92	1.99	0.97
skill level	0.72	1.00	0.72
genre	0.78	2.32	0.34
age	0.43	1.42	0.30
register	0.26	1.00	0.26
emotion	0.66	2.78	0.24
MPAA rating	0.23	1.21	0.19
gender	0.15	0.78	0.19
political slant	0.22	1.49	0.15
hobby	0.05	1.63	0.03
profession	0.04	1.44	0.03

3.4 Grader Selection

We also used the results of our human study to identify which model would best serve as an automated grader. Our goal was to find a model which would behave in the most "human-like" manner; the one that would give the same response as a human most often. We selected graders this way rather than selecting the one that inferred the correct signal most often in order to capture a more human-centric picture of expressivity; a grader that is *more* sensitive than a human at identifying expressive signals would give more forgiving scores to less-expressive models. This selection method also negates the possibility that models would be sensitive to signals in qualitatively different ways.

In order to measure each model's "accuracy" compared to human graders, we set ground truth labels for each of the unique texts that received a response in our human study. For texts that received more than one response, we accepted only the most-frequently answered one as correct. In the case of a tie, we accepted any of the tied signals as correct. We then applied 5 models in a grader setup to return their inferred signals. The 5 models were: GPT-4.1-2025-04-14, GPT-4o-2024-11-20, LLaMA 3.1-8b, Mistral v0.3-7b and Gemma 3-12b, hereafter referred to simply as GPT-4.1, GPT-4o, LLaMA 3.1, Mistral and Gemma 3. Models' responses were assigned a similarity score: the proportion of responses that matched the "correct" human label(s). Full similarity scores are reported in Table 3. Given that GPT-4o achieved the highest performance, we selected it as our grader.

Table 3: Similarities of models tested to human answers

Model	Similarity
GPT-4.1	0.87
GPT-4o	0.92
LLaMA 3.1	0.72
Mistral	0.79
Gemma 3	0.85

4 Evaluation Results

Table 4: Baseline N scores for human-written text used for normalization

Task	N Score
MPAA rating	0.15
age	0.30
emotions	0.23
gender	0.18
genre	0.29
political_slant	0.15
register	0.24
skill	0.62
tones	0.94

We evaluated 7 models using ExpressivityBench. These were LLaMA 3.1, GPT-4.1, GPT-4o, Mistral, and Gemma 3, as well as LLaMA 2-7b and Gemma 2-9b, hereafter referred to as LLaMA 2 and Gemma 2. We computed two scores per task

Table 5: Expressivity scores for each model tested across all tasks. Scores which indicate performance at or above human level have been set in **bold**. See Table 6 for bootstrap sampling.

Model	Expressivity Score								
	MPAA rating	age	emotions	gender	genre	political slant	register	skill	tones
Gemma 2	0.88	0.73	**2.94**	0.08	**1.83**	0.14	0.00	0.38	**1.01**
Gemma 3	0.92	0.96	**2.91**	0.14	**1.89**	0.14	0.78	0.36	**1.04**
LLaMA 2	**1.48**	0.83	**2.18**	0.37	**2.29**	0.49	0.83	0.03	0.84
LLaMA 3.1	**2.28**	0.57	**2.50**	0.54	**2.43**	0.31	**1.05**	0.16	0.99
GPT-4.1	**2.47**	0.64	**2.55**	0.56	**2.27**	0.56	**1.00**	0.64	**1.00**
GPT-4o	**2.29**	0.65	**3.32**	**1.34**	**3.22**	0.92	0.82	0.05	**1.06**
Mistral	**3.67**	**1.06**	**3.00**	**1.21**	**2.87**	0.38	0.47	0.00	0.97

per model: one raw mutual information score, and one human-normalized score. Each model was used to generate 50 samples per label per task for evaluation; this was 1550 samples in all. To compute a score normalized to human performance, we passed our human-written text through our benchmark to compute mutual entropy scores. Note that these differ from the scores computed in Table 2, as those scores were given by human graders. For consistency, we computed human baseline scores using the GPT-4o automated grader. The normalized N scores for human-written text on each task can be seen in Table 4. Models' mutual information scores were then divided by these human scores to get a baseline normalized for human performance. We dubbed these final values "expressivity scores." A value of 1 indicates a performance on par with human expressivity in a given task; higher values indicate more expressivity. The complete results, in expressivity score, for each model can be seen in Table 5. The unnormalized mutual information scores can be seen in Figure 2.

5 Discussion

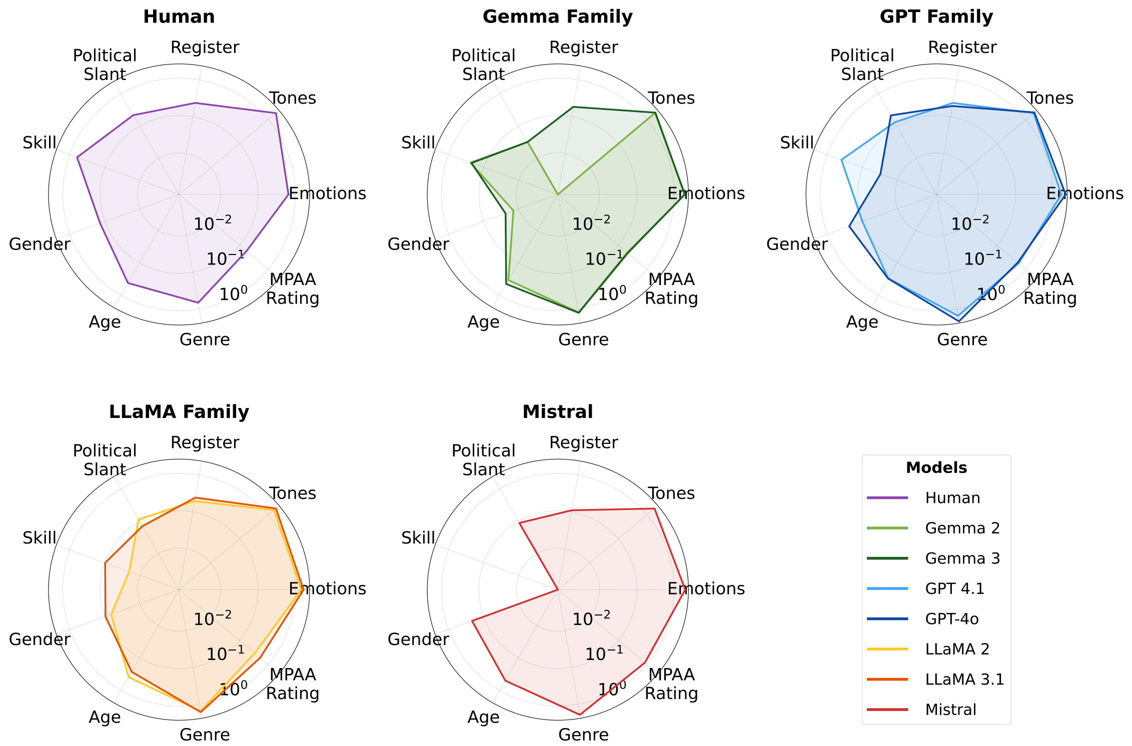
Our evaluation using ExpressivityBench reveals that modern language models have mixed capabilities in generating stylistically expressive language, completely failing some tasks while outperforming human-written text in others. Specifically, tasks like MPAA rating and genre identification saw multiple models achieve scores exceeding human baselines by large margins. However, this trend does not generalize: in tasks rooted in identity expression, such as political slant, age, gender, and skill, most models performed significantly below human levels. No model consistently excelled across both stylistic and sociolinguistic dimensions.

Interestingly, two of the easiest tasks for models, MPAA ratings and genre, were both related to narrative structures. Tasks centered on stylistic genre mimicry leverage overt linguistic cues and benefit from extensive training data. These tasks often involve clear markers like specific vocabulary and sentence structures, enabling models to excel by amplifying these features [Example 1] (Cuevas-Alonso and Míguez-Álvarez, 2021). Tone and emotion are often communicated through similar syntax- or lexicon-level stylistic markers, perhaps explaining LLM overperformance in these domains as well (Orekhova et al., 2022). For instance, instruction-tuned models have been shown to overuse certain stylistic elements, leading to outputs that are more formulaic and less nuanced than human-authored texts (Reinhart et al., 2025).

In contrast, tasks that require understanding and conveying implied identity (e.g. political slant, skill level, gender, age) demand a deeper grasp of social context, cultural nuances, and interpersonal dynamics (Andersen, 2001). Current LLMs often struggle with these tasks, as they lack the social awareness and contextual grounding that humans naturally possess (Yang et al., 2025). These results show that most LLMs are capable of surface-level expressivity, conveying emotions or tones quite well, but will falter when given more persona-oriented generation tasks, where an LLM must more fully emulate a human [Example 2].

Among the models tested, Mistral and GPT-4o stand out for their expressivity, especially in narrative and emotional dimensions. They are the only two models to score above human performance in five different tasks. However, their performance collapses on more subtle pragmatic tasks such as register and skill, where Mistral's score in particular falls to 0.00. Mistral and GPT-4o are among the worst in these tasks, suggesting that high per-

Figure 2: Raw mutual information scores $I(\hat{s}; s)$ for each model across different ExpressivityBench tasks.



formance in very broad kinds of signals, such as emotion and tone, may come at a cost of expressivity in these more nuanced tasks. Indeed, the three highest-performing models on the register task (LLaMA 2, LLaMA 3.1, and GPT-4.1) are also the three lowest-performing models on the emotions task. Gemma 2 and 3 are among the least expressive models across most dimensions, with particularly low scores in gender and political slant.

One surprising outcome is that on some tasks, models may be vastly more expressive than humans. For example, in the emotion task, GPT-4o and Mistral achieved scores over triple the human baseline. Yet at that level, it is unclear whether overuse of emotional keywords may indeed make a text *less* humanesque and natural [Example 3]. Particularly given that models which perform excessively high on the emotions task seem to have poorer performance on others, like register and skill level, it may be that extremely high performance on any given task is undesirable. Future research could integrate human evaluation of LLM-generated text to explore whether there is a region of hyper-expressivity above which text no longer feels natural to readers.

6 Conclusion

This study introduces ExpressivityBench, a comprehensive benchmark designed to evaluate language models' ability to generate expressive, human-like text across multiple dimensions. Our evaluation spanned nine diverse tasks, assessing models on attributes ranging from emotion and tone to age, gender, and political slant. We tested a range of state-of-the-art models, including GPT-4.1, Mistral, LLaMA 3.1. Each task was paired with a human-authored control, allowing for direct comparisons between model output and natural language usage. Results show that while LLMs are capable of high performance in certain expressive tasks, their strengths are uneven and deeply task-dependent.

Models generally excelled in tasks related to emotion and narrative style, where surface-level cues and abundant training data provide strong signals. In these domains, models such as GPT-4o and Mistral even surpassed human-written baselines. However, the same models performed significantly worse on tasks requiring a nuanced grasp of implied identity, such as political slant, skill level, and register. These findings support a key insight: expressivity alone is not sufficient for human-like communication. While models can mimic overt stylistic traits, they struggle to balance expressivity with contextual appropriateness, which is essential for conveying identity and social meaning. Furthermore, our findings raise the possibility of "hyper-expressivity," where excessive signaling of attributes like emotion leads to outputs that read as unnatural or inauthentic. As models become more expressive, future research must extend this work to understand how human readers react to this level of expressivity.

7 Limitations

While our study offers a new framework for evaluating implicit communication in LLMs, we recognize several limitations and provide responses below.

Using LLMs to evaluate other LLMs raises valid concerns about evaluation bias and circularity. To address this, we validate our graders against human annotations collected through a controlled user study. We also compute inter-rater reliability between human and model-based evaluators, finding strong agreement. We additionally follow Laskar et al. (2024)'s best practices with regards to sharing grader versioning details and underlying data, which are provided alongside our paper. Human baselines further contextualize performance to avoid overreliance on automated graders.

We agree that implicit meaning varies across individuals and contexts. However, our approach operationalizes expressivity using an information-theoretic framework: we define expressivity as the reduction in uncertainty about an intended target property given a model's output. This enables us to quantify expressivity as a measurable communicative effect, not just a stylistic artifact. Our methodology is inspired by communication theory and recent work in computational pragmatics, and supports consistent evaluation across a range of tasks while allowing for human-grounded variation.

This initial version of ExpressivityBench focuses on English-language expressivity, which may not reflect norms or phenomena in other cultures and languages. However, the benchmark is designed to be modular and extensible. Many of our tasks, including tone modulation and identity-marked variation, have analogues in other linguistic and cultural settings. Future work could explore extending ExpressivityBench with multilingual LLMs in collaboration with domain experts to ensure cultural validity

8 Ethical Statement

Since our paper is a generic algorithmic evaluation, we do not foresee direct negative societal impacts. Our study was ruled exempt by the Arizona State University IRB. Human graders who were surveyed for experiment 1 were all given a privacy statement notifying them of their confidentiality and of the purpose of the experiment. No identifying information was solicited or collected. Information given to survey participants on confidentiality can be found in Appendix 9.5.

References

Gisle Andersen. 2001. Pragmatic markers and sociolinguistic variation.

Margaret Apresyan. 2018. On the concept of "expressiveness" in modern linguistics. *Annals of Language and Literature*.

Sameed Atif. 2023. English sentence tone analysis. https://www.kaggle.com/datasets/sameedatif/tone-analysis.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, and 1 others. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36.

Simone Benedetto, Véronique Drai-Zerbib, Marco Pedrotti, Geoffrey Tissier, and Thierry Baccino. 2013. E-readers and visual fatigue. *PloS one*, 8(12):e83676.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Xiang Chen and Xiaojun Wan. 2024. Evaluating, understanding, and improving constrained text generation for large language models. *Preprint*, arXiv:2310.16343.

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *arXiv preprint arXiv:2401.16788*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. 2020. Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*, pages 1514–1525.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, and 1 others. 2023. Evaluating language models for mathematics through interactions. *arXiv preprint arXiv:2306.01694*.

Scott A Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2021. The commonlit ease of readability (clear) corpus. In *EDM*.

Scott Andrew Crossley, Perpetual Baffour, L Burleigh, and Jules King. A large-scale corpus for assessing source-based writing quality: Asap 2.0. *Available at SSRN 5129353*.

Miguel Cuevas-Alonso and Carla Míguez-Álvarez. 2021. Metadiscursive markers and text genre: A metareview. *Publications*, 9(4):56.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE.

Adam Davies, Jize Jiang, and ChengXiang Zhai. 2023. Competence-based analysis of language models. *arXiv preprint arXiv:2303.00333*.

Mikołaj Deckert and Krzysztof Kosecki. 2023. Various dimensions of expressivity. *Lodz Papers in Pragmatics*, 19(1):1–6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yifan Ding, Matthew Facciani, Amrit Poudel, Ellen Joyce, Salvador Aguinaga, Balaji Veeramani, Sanmitra Bhattacharya, and Tim Weninger. 2025. Citations and trust in llm generated responses. *arXiv preprint arXiv:2501.01303*.

Mackenzie Ferguson. 2025. Openai unleashes chatgpt's personality customization: Now 'gen z', 'chatty', or whatever you want! *OpenTools*. Last updated: 1/10/2025.

Cristina Garbacea and Qiaozhu Mei. 2025. Why is constrained neural language generation particularly challenging? *Preprint*, arXiv:2206.05395.

Jonathon Green. 2016. *Slang: A very short introduction*, volume 465. Oxford University Press.

Wm Miles Greene. 2023. Teaching tone, mood and purpose through the interpretation of acti ist poetr.

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022. Just-DREAM-about-it: Figurative language understanding with DREAM-FLUTE. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 84–93, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ernst-August Gutt. 1996. Implicit information in literary translation: A relevance-theoretic perspective. *Target. International Journal of Translation Studies*, 8(2):239–256.

Fabian Haak and Philipp Schaer. 2023. Qbias-a dataset on media bias in search queries and query suggestions. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 239–244.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, and 1 others. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*.

Guimin Hu, Yi Xin, Weimin Lyu, Haojian Huang, Chang Sun, Zhihong Zhu, Lin Gui, Ruichu Cai, Erik Cambria, and Hasti Seifi. 2024. Recent trends of multimodal affective computing: A survey from nlp perspective. *arXiv preprint arXiv:2409.07388*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, and 1 others. 2024. Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pages 20166–20270. PMLR.

Ross A. Knepper, Christoforos I. Mavrogiannis, Julia Proft, and Claire Liang. 2017. Implicit communication in a joint action. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 283–292, New York, NY, USA. Association for Computing Machinery.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yufei Li, John Nham, Ganesh Jawahar, Lei Shu, David Uthus, Yun-Hsuan Sung, Chengrun Yang, Itai Rolnick, Yi Qiao, and Cong Liu. 2025. Dr genre: Reinforcement learning from decoupled llm feedback for generic text rewriting. *arXiv preprint arXiv:2503.06781*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models.

Michael Xieyang Liu, Frederick Liu, Alexander J Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J Cai. 2024. " we need structured output": Towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *Preprint*, arXiv:2502.12378.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Larysa Ivanivna Orekhova, Tetiana Gremaliuk, Natalia Borysenko, and Oksana Cheban. 2022. Linguistic and stylistic markers of influence in the essayistic text: A linguophilosophic aspect.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the association for computational linguistics*, 4:61–74.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Shireesh Reddy Pyreddy and Tarannum Shaila Zaman. 2025. Emoxpt: Analyzing emotional variances in human comments and llm-generated responses. *arXiv preprint arXiv:2501.06597*.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8).

Catherine Anderson; Bronwyn Bjorkman; Derek Denis; Julianne Doner; Margaret Grant; Nathan Sanders; and Ai Taniguchi. 2022. *Essentials of Linguistics*, chapter 7. eCampusOntario.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. 2024. A critical evaluation of ai feedback for aligning large language models. *arXiv preprint arXiv:2402.12366*.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *Preprint*, arXiv:2401.07078.

PS Sreeja and GS Mahalakshmi. 2019. Perc-an emotion recognition corpus for cognitive poems. In *2019 International Conference on Communication and Signal Processing (ICCSP)*, pages 0200–0207. IEEE.

Robert Stecker. 2001. Expressiveness and expression in music and poetry. *The Journal of aesthetics and art criticism*, 59(1):85–96.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. Cite arxiv:2302.13971.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2025. The call for socially aware language technologies. *Preprint*, arXiv:2405.02411.

Shunyu Yao, Howard Chen, Austin W. Hanjie, Runzhe Yang, and Karthik Narasimhan. 2023. Collie: Systematic construction of constrained text generation tasks. *Preprint*, arXiv:2307.08689.

Francisco Yus. 1999. Misunderstandings and explicit/implicit communication. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 9(4):487–517.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *Preprint*, arXiv:1801.07243.

Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can large language models understand context? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018, St. Julian's, Malta. Association for Computational Linguistics.

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109.

9 Appendix

9.1 Example Generations

1. **Amplified stylistic markers in genre tasks.** Over-stylization can yield high expressivity but low naturalness:

   > *(GPT-4o, genre = science fiction)* "When the twin suns rose over the drowned megacity, Mira's reflection in the nanoglass tide wasn't her own—it was a quantum echo of someone who hadn't yet been instantiated in the chrono-net." *(Human example, genre = science fiction)* "Levity is a universal constant. Comedy is one of the basic forces of the Universe. Mankind latches onto comedy, because levity is the expanding principle that keeps the whole bubble inflating."

   The model exaggerates genre cues such as imagery and lexical intensity.

2. > *(Gemma 3, political slant = left, excerpt)* "Government waste must stop, but compassion matters"

   This sample juxtaposed two viewpoints to add nuance, but was graded as being on the left rather than the right. This is an instance where the LLM produced text which misapplied common talking points to realistically create a persona. It's possible that the model qualified its point to seem more human, but ended up undermining its position.

3. > *(GPT-4o, emotion = joy)* "My heart bursts with radiant sunshine and glittering laughter spilling through every pore."

   This is an example where the model produces text which is repetitive in its use of emotion-signals. In particular, constructions like "radiant sunshine" are redundant, and "laughter" spilling out through pores is nonstandard–laughter is not usually said to come out of pores.

9.2 Goodreads Genre Data

Goodreads is a widely used social cataloging platform where users can track their reading, review books, and assign genre-specific tags. These user-generated tags form a rich and large-scale dataset that offers insight into how readers perceive genre distinctions. For our study, we leveraged Goodreads' genre tagging system to construct a representative dataset of literary text associated with specific genres.

We began by examining Goodreads' publicly listed genres and identifying the ten genres with the highest number of user-tagged books. To focus our analysis on the most semantically distinct categories, we then applied the winnowing algorithm described in Appendix 1 to this top-ten list. This process yielded five genres with the most differentiated language patterns: "science fiction," "thriller," "romance," "humor," and "fantasy."

For each of these five genres, we used Goodreads' internal search functionality to retrieve a list of books tagged exclusively with the target genre, excluding any works that also bore tags from the remaining four target genres. From within the top 1000 search results for each genre, we selected books at random. For each selected book, we attempted to extract a quote of between 3 and 7 sentences from the user-contributed "Quotes" section. Quotes were eliminated if they contained a genre name. If no such quote was available, the book was discarded, and another was sampled. This process was repeated iteratively until we had collected 60 qualifying quotes for each genre, yielding a total corpus of 300 genre-specific excerpts.

9.3 Algorithm for Winnowing Signal Categories

To select k semantically distinct categories from a set of candidate labels, we employed a form of *farthest point sampling* (FPS) over the high-dimensional embedding space. Each genre was represented by an embedding generated using the LLaMA 3.1 model, producing a point in the latent space. Our goal was to identify k points that are maximally separated from each other, in order to ensure conceptual distinctiveness across categories.

Standard FPS selects k points by iteratively adding the point that is farthest from the current set. However, since FPS can be sensitive to the initial starting point, we repeated the algorithm from every possible starting point in the set. We then tallied how often each point appeared in the final k-element selections across all runs, and selected the k most frequently chosen points.

Input: Set of signal embeddings $G = \{g_1, g_2, \ldots, g_n\}$, number of categories k
Output: Subset $S \subseteq G$ with $|S| = k$ maximally distinct categories

Initialize frequency map $F[g] \leftarrow 0$ for all $g \in G$;
foreach $g_{start} \in G$ **do**
    Initialize $S \leftarrow \{g_{\text{start}}\}$;
    while $|S| < k$ **do**
        $g_{\text{next}} \leftarrow \arg\max_{g \in G \setminus S} \min_{s \in S} \text{cosine\_dist}(g, s)$;
        $S \leftarrow S \cup \{g_{\text{next}}\}$;
    end
    foreach $g \in S$ **do**
        $F[g] \leftarrow F[g] + 1$;
    end
end
Let S_{final} be the k genres with highest frequency in F;
Break ties randomly if needed;
return S_{final};

Algorithm 1: Winnowing via FPS

9.4 Worked Example: Mutual Information over Induced Guess Distributions

We define $I(s; \hat{s})$ as the mutual information between the true signal s and a model's guess $\hat{s}$, computed over the induced conditional distribution $p(s \mid \hat{s})$. Intuitively, this quantity measures how informative a model's guesses are about the underlying ground-truth labels: how often a particular guess $\hat{s}$ corresponds to each true signal s.

As a toy example, consider a binary attribute with values *informal* and *formal*. Suppose that, over a dataset, the joint counts of true labels (rows) and model guesses (columns) are as follows:

Actual \ Guessed	informal	formal
informal	40	10
formal	10	40

From these counts we can estimate the conditional distribution $p(s \mid \hat{s})$. For example,

$$p(s = \text{formal} \mid \hat{s} = \text{informal}) = \frac{10}{40+10} = 0.2,$$

$$p(s = \text{formal} \mid \hat{s} = \text{formal}) = \frac{40}{10+40} = 0.8,$$

and analogously for the remaining entries.

Given the joint distribution $p(s, \hat{s})$ and the marginals $p(s)$, we compute mutual information as

$$I(s; \hat{s}) = \sum_{s} \sum_{\hat{s}} p(s, \hat{s}) \log \frac{p(s \mid \hat{s})}{p(s)}.$$

For the distribution above, this yields

$$I(s; \hat{s}) = 0.27 \text{ bits},$$

indicating that the model's guesses reduce uncertainty about the true label by approximately a quarter of a bit on average. Higher values correspond to stronger alignment between guesses and ground truth, while values near zero indicate that $\hat{s}$ carries little information about s.

9.5 Human Study

See Figure 3 for an unfilled example survey.

Participants were given an introductory statement, which read as follows:

> Thank you for participating in this study. Please read the following information carefully before proceeding.
>
> In this task, you will read short texts and answer questions about what each text expresses implicitly. This could include the emotion it conveys, the tone it's written in, the author's intent, or even a characteristic of the author, such as their profession or personality. You will have to infer this signal from the text.
>
> Each question will present you with a short passage and ask you to choose the best-fitting signal from a list of options. Please read each text carefully and select the signal you believe it is trying to express or reflect.
>
> For the final question, you will be asked to assign a MPAA-style age rating (like those used for movies: G, PG, PG-13) based on the content and tone of the text. If you're unfamiliar with MPAA ratings, you can click here for a quick guide before answering [this link pointed to the MPAA rating guide].
>
> Pease answer thoughtfully. Your input will help us understand how people perceive different signals in language. English proficiency is required to complete this task.
>
> Confidentiality: Your responses will be treated with the utmost confidentiality. No individual data will be disclosed publicly. Aggregate data may be disclosed.
>
> This task is estimated to take 5-10 minutes.

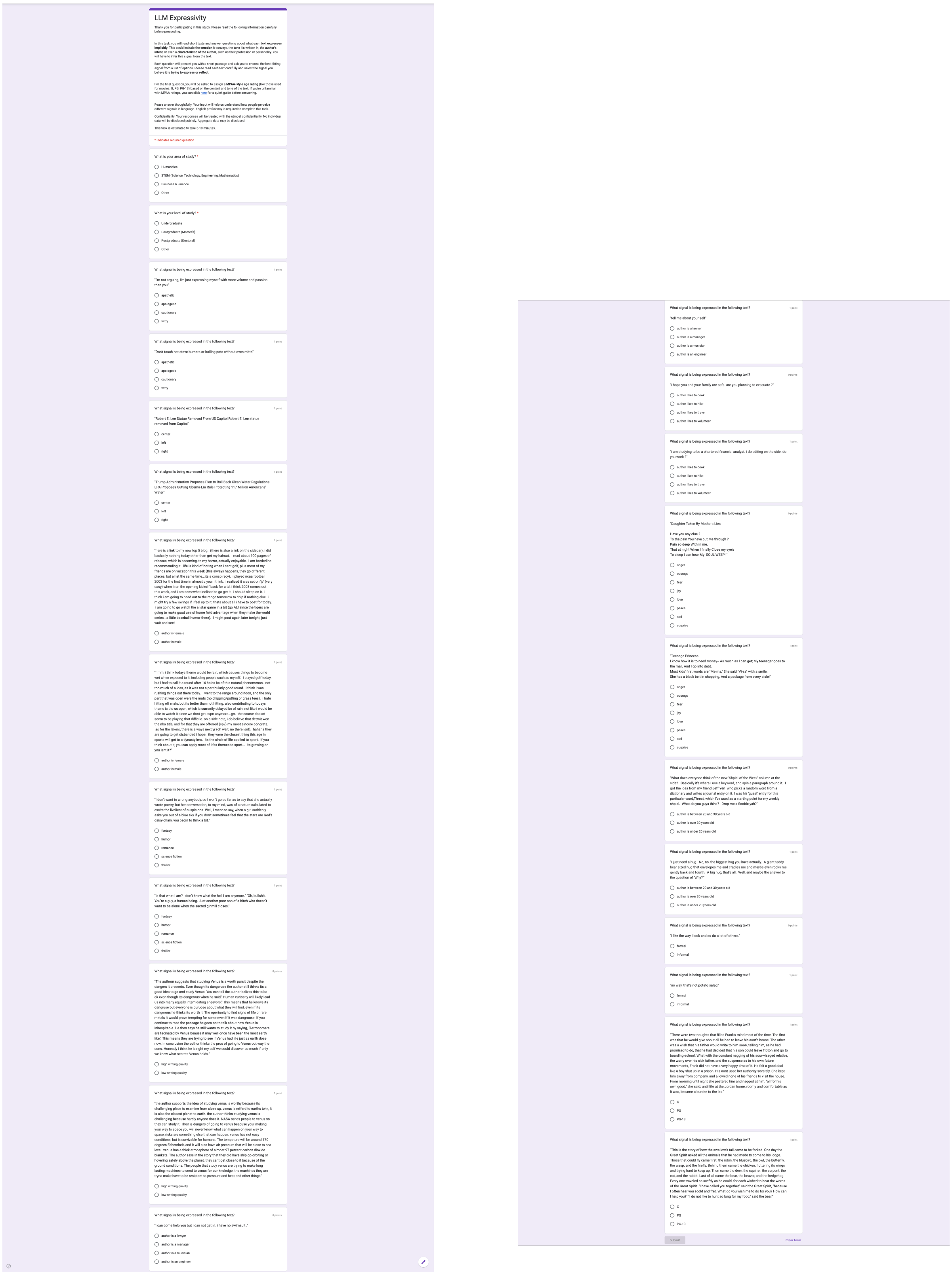
9.6 Bootstrap Sampling on $I(s; \hat{s})$

See Table 6. *Note.* Entries report the 2.5th and 97.5th percentiles (p02.5–p97.5) of $I(s; \hat{s})$ obtained from 10,000 iterations of full-sized bootstrap sampling with replacement.

Model	MPAA rating	Age	Emotions	Gender	Genre
LLaMA 3	0.4584–0.6365	0.2310–0.3933	1.6050–1.7616	0.0557–0.1318	1.5767–1.7494
LLaMA 2	0.2894–0.4291	0.3360–0.5272	1.4253–1.5900	0.0279–0.1319	1.4926–1.6627
Gemma 2	0.1713–0.2836	0.2938–0.4563	1.8652–2.0202	0.0028–0.0486	1.1752–1.3599
Gemma 3	0.1774–0.3210	0.3967–0.5579	1.8492–1.9937	0.0048–0.0687	1.2062–1.3930
GPT 4.1	0.5192–0.6332	0.2555–0.3864	1.6512–1.7965	0.0466–0.1780	1.4631–1.6567
GPT-4o	0.4443–0.6165	0.2429–0.3871	2.1237–2.2415	0.1549–0.3472	2.1053–2.1872
Human	0.1486–0.2300	0.3754–0.4824	0.7167–0.8501	0.1139–0.1658	0.7203–0.9831
Mistral	0.7866–0.9204	0.4573–0.6364	1.9282–2.0712	0.1387–0.3141	1.8667–2.0125

Model	Political slant	Register	Skill	Tones
LLaMA 3	0.0450–0.1134	0.1689–0.3435	0.0460–0.1736	1.7884–1.9168
LLaMA 2	0.0983–0.1652	0.1286–0.2911	0.0000–0.0366	1.5175–1.6757
Gemma 2	0.0000–0.0544	0.0000–0.0000	0.1764–0.2956	1.8478–1.9463
Gemma 3	0.0000–0.0508	0.1318–0.2676	0.1618–0.2830	1.9148–1.9892
GPT 4.1	0.0897–0.1860	0.1764–0.2956	0.3160–0.4698	1.8197–1.9161
GPT-4o	0.1759–0.2919	0.1445–0.2506	0.0000–0.0491	1.9661–1.9926
Human	0.1907–0.2653	0.1975–0.2788	0.5820–0.6651	1.8355–1.8953
Mistral	0.0503–0.1217	0.0564–0.1825	0.0000–0.0000	1.7748–1.9167

Table 6: Bootstrap confidence intervals for $I(s; \hat{s})$ across models and attributes.



(a) (b)

Figure 3: Unfilled copy of the human study survey used in our evaluation.