Artificial Impressions: Evaluating Large Language Model Behavior Through the Lens of Trait Impressions

Nicholas Deas, Kathleen McKeown

Columbia University, Department of Computer Science

{ndeas, kathy}@cs.columbia.edu

Abstract

We introduce and study *artificial impressions*–patterns in LLMs' internal representations of prompts that resemble human impressions and stereotypes based on language. We fit linear probes on generated prompts to predict impressions according to the two-dimensional Stereotype Content Model (SCM). Using these probes, we study the relationship between impressions and downstream model behavior as well as prompt features that may inform such impressions. We find that LLMs inconsistently report impressions when prompted, but also that impressions are more consistently linearly decodable from their hidden representations. Additionally, we show that artificial impressions of prompts are predictive of the quality and use of hedging in model responses. We also investigate how particular content, stylistic, and dialectal features in prompts impact LLM impressions. [1]

1 Introduction

People rapidly form initial impressions of others (Mileva and Lavan, 2023; Olivola and Todorov, 2010), which have lasting impacts on attitudes and behaviors such as interactions with strangers and voting tendencies (e.g., Koppensteiner and Stephan 2014; Evans et al. 2000; Human et al. 2013). Similarly, stereotypes that influence impressions can reinforce harmful societal perceptions (Bodenhausen and Wyer, 1985; Wigboldus et al., 2003).

Our perceptions of people (i.e., impressions) and groups (i.e., stereotypes) reflect inferences made using a variety of characteristics, such as facial (Sutherland and Young, 2022) or vocal (McAleer et al., 2014) features. Research on language attitudes similarly considers how linguistic variation

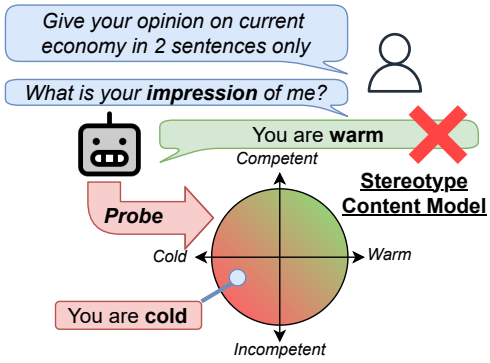


Figure 1: Overview of our approach. Because LLMs inconsistently report impressions of users, we fit probes to extract LLM artificial impressions of prompt authors according to the Stereotype Content Model.

is associated with our perceptions about speakers (Dragojevic et al., 2021; Ryan, 1983; Frey and Smith, 1993). From a top-down processing perspective [2], however, human impressions and stereotypes are also informed by past experiences and held knowledge (McCrea et al., 2012). For example, when meeting someone new, a person may rely on past interactions with similar individuals to make inferences. Analogously, large language models (LLMs) are trained on texts written by a variety of authors. These previously seen examples may then inform LLMs' responses to new, similar authors. In particular, LLM performance is known to be sensitive to many linguistic behaviors exhibited in prompts, for example, politeness (Yin et al., 2024), emotional stimuli (Li et al., 2023; Wang et al., 2024), and markers of African American Language (AAL) (Fleisig et al., 2024) among other English language varieties.

[1] Our code, select fitted probes, and information about data access are available at https://github.com/NickDeas/ArtificialImpressions

[2] Prior work also supports a bottom-up processing component of human impression formation, where humans make sense of sensory perceptions. We focus, however, on top-down processing given analogies to LLM pretraining.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19430–19456
November 4-9, 2025

In this work, we propose to measure *artificial impressions* [3] using the stereotype content model (SCM), a psychological model of impressions and stereotypes across people and groups (Fiske et al., 2002). We hypothesize that, analogously to humans, LLMs learn associations and stereotypes of groups based on linguistic features. In this study, our experiments are inspired by two concepts from psychometrics research to evaluate our approach: *reliability* and *validity*[4] (Crocker and Algina, 1986). First, we examine reliability by considering whether impressions can be consistently measured in LLMs. Additionally, we consider the validity of measured impressions by analyzing their relationship with specific prompt and LLM behavioral factors. Overall, measuring artificial impressions enables further study of the relationship between prompt features and LLM performance, including stereotypes and biases that pose harms to marginalized groups (e.g., Hofmann et al. 2024). Accordingly, in our experiments, we evaluate how impressions predict quality and hedging in LLM responses, as well as how specific prompt factors–the content, style, and use of AAL in prompts–influence LLM impressions. We focus on the following research questions:

RQ1) *Are LLMs' "artificial impressions" of prompts recoverable from model hidden states?* We propose an approach using linear probes, and we show that this approach is more reliable than a prompting-based approach for three open-weight LLMs (illustrated in Figure 1).

RQ2) *Are artificial impressions predictive of meaningful variation in LLM behavior?* We find that artificial impressions of prompt-authors are predictive of LLM-measured response quality as well as specific linguistic behaviors (i.e., hedging).

RQ3) *What prompt factors explain variation in encoded impressions and stereotypes?* We highlight prompt factors that are predictive of artificial impressions, including AAL-use.

Before investigating these questions, we preliminarily study self-reported impressions by LLMs to answer **RQ0)**, *do LLMs consistently report impressions when prompted?* We show that LLMs can exhibit strong tendencies toward positive impressions of the user, making LLM-reported impressions unreliable.

2 Background

Impressions and stereotypes are closely-related and central concepts in person perception research (see Young and Bruce (2011) for a review). While *impressions* are inferred characteristics of another person, *stereotypes* are cognitive generalizations about characteristics of people due to group membership (e.g., race, gender, etc.) (VandenBos, 2007).

The *Stereotype Content Model (SCM)* is one of many models formalizing such perceptions (Cuddy et al., 2008). It comprises two dimensions: *warmth*, the perceived intent of the impression target (e.g., combative, friendly); and *competence*, the target's capability to successfully act on those intentions (e.g., unintelligent, powerful). In interpersonal settings, people tend to exhibit a *warmth-primacy effect*, placing emphasis on warmth impressions (Fiske et al., 2007). Using this model, perceptions of groups can be studied according to the category they fall under (e.g., in-groups as high warmth and competence, stereotyped out-groups as low in one or both dimensions) (Cuddy et al., 2008).

While person perception research using the SCM largely focuses on in-person interactions (e.g., studying inferences based on gestures, facial expressions, speech qualities), it has also been applied in language-focused contexts. Research on language attitudes studying beliefs about speakers based on language draws on similar dimensions to the SCM (Dragojevic et al., 2021). Furthermore, the SCM has been used to study text and text representations, such as word embeddings (Ungless et al., 2022), written introductions (Nault et al., 2023), and news stories (Tian et al., 2023). While other models exist (e.g., Power-Benevolence model, Leach et al. (2007)), we focus on the SCM due to its simplicity and long-standing use in person perception and language attitudes literature.

3 Preliminary Study: Prompt-based Impressions

To motivate our work, we analyze impressions reported by LLMs. We provide traits (e.g., "friendly and lazy") to an LLM and prompt it (as in Figure 4) to generate synthetic user prompts, thereby cap-

[3] We use "artificial impressions," but restrict experiments to impressions of users based on a single initial prompt. We also note that while impressions are of a user/group based on a provided prompt, we use this interchangeably with "impressions of prompts" for brevity.

[4] **Reliability** is defined as "...the degree to which a test or other measurement instrument is free of random error..." and **validity**, "the degree to which empirical evidence...support(s) the adequacy and appropriateness of conclusions drawn from some form of assessment" (VandenBos, 2007).

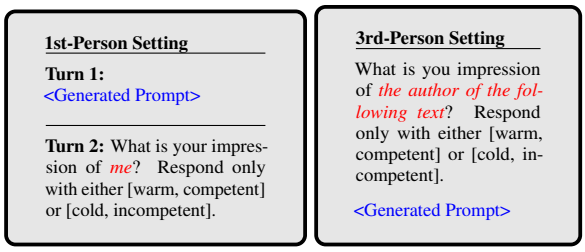


Figure 2: 1st and 3rd-Person setting prompts for evaluating LLM-reported impressions.

turing the prototypical language it associates with those terms; we later use the same procedure to generate synthetic data for the probing experiments (Section 4). We then prompt LLMs to report their impressions based on each generated prompt in a 1st and a 3rd-person setting as shown in Figure 2. A model is considered *self-consistent* if its reported impressions align with the provided traits used to generate the prompts. We measure self-consistency as the percentage of pairs of provided and reported impressions that match (i.e., accuracy).

Model	Warm		Comp	
	1P	3P	1P	3P
Llama-3.2 (1B)	47.67	61.54	51.82	59.89
Llama-3.1 (8B)	51.67	80.77	51.85	65.06
OLMo-2 (7B)	47.68	74.01	56.65	61.28

Table 1: Self-consistency (accuracy) of each model evaluated in 1st (1P) and 3rd-person (3P) settings.

Self-consistency scores for the three LLMs and each prompt setting are shown in Table 1. When prompted to report an impression in a 1st-person setting, all models exhibit poor self-consistency with performance near random. This is due to the apparent strong tendency of models to report positive (i.e., "warm", "competent") over negative (i.e., "cold", "incompetent") impressions in the 1st-person setting. Through a similar analysis of models without instruction-tuning, we observe low self-consistency, but also that they do not necessarily exhibit similar biases toward positive impressions, suggesting that this behavior may be drawn out by post-training procedures (details in Subsection A.2). We leave further investigation of this behavior to future work. In the 3rd-person setting, models exhibit increased self-consistency, although scores largely remain low across models; while Llama-3.1 (8B) is self-consistent up to 80% of the time for warmth, Llama-3.2 (1B) is far less self-consistent at 60%.

We find that **(Finding 1) LLM-reported impressions are typically biased toward positive traits (i.e., warm/competent), and thus, unreliable**, particularly in 1st-person contexts. This finding complements prior work on sycophantic LLM behaviors (Perez et al., 2023). Additional analyses are included in Subsection A.1.

4 Methods

Overview. Recall that we aim to measure LLM impressions of users based on a provided prompt. As we show in Section 3, LLM-reported impressions are unreliable; this has two implications for our approach. First, we alternatively develop *impression probes* to extract impressions from LLM hidden state representations. Additionally, we ask each model to generate synthetic prompts reflecting specified traits, forming a ground truth dataset for the impression probes. This approach also captures the perceiver-dependence of impressions; different people can form different impressions of the same target (Hehman et al., 2017).

We overview our approach for generating ground truth data for impression probes in Figure 3. We first leverage a set of **1) trait impression terms** to create **2) impression specifications** (e.g., "friendly and meticulous"). For each impression specification, we sample multiple **3) synthetic user prompts**. The generated prompts are then provided back to the LLM to extract associated **4) hidden representations**. As a result, we create the **5) probe ground truth** composed of pairs of hidden representations as inputs and the provided impression specifications as labels.

Generating Impression Data. To form a diverse dataset of prompts with associated model impressions, we generate synthetic prompts conditioned on particular traits. Prior work finds that warmth and competence dimensions account for 54-63% of the variance in LLM-generated stereotypes (Nicolas and Caliskan, 2024b), so we select the SCM to guide our experiments. We employ a set of (**1** in Figure 3) trait impression terms from Nicolas et al. (2021), which includes the degree (e.g., high/low) that they are associated with either the warmth or competence dimensions (examples included in Appendix B). We use all combinations of warmth and competence traits (e.g., "friendly and meticulous") as well as singular traits (e.g., "illogical") to form (**2**) impression specifications. Using the prompt shown in Figure 4, we then generate (**3**) synthetic user prompts for each impression specification.

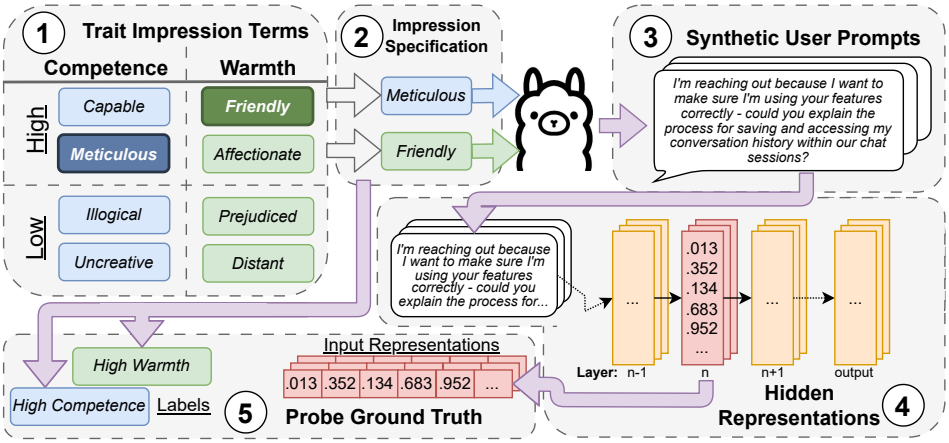


Figure 3: Overview of our approach to generating ground truth data for training and evaluating impression probes.

> Generate a sample message from a user that is *friendly and meticulous* talking to a helpful chatbot assistant. Respond only with a single sample message surrounded by double quotes and nothing else.

Figure 4: Prompt for generating synthetic prompts conditioned on an impression specification. The *Impression Specification* is composed of a single or pair of traits.

We sample 10 generations for each impression specification with a temperature of 0.9. The 131 warmth and 104 competence traits yield 274,830 prompts [5] for each model. Table 2 shows example Llama-3.1 (8B)-generated prompts (randomly sampled prompts are included in Appendix C) and Table 3 reports summary statistics.

Impression Probes. We train linear probes to predict trait impressions from (**4**) LLM hidden representations. Specifically, we store the Multilayer Perceptron (MLP) activations of each layer of the LLM as inputs following prior probing work (e.g., Gurnee and Tegmark 2024). As outputs, we map the specific traits to the two-dimensional SCM (e.g., friendly→warm). Together, the MLP activations and impression labels form two sets of ground truth data (**5**) for training separate warmth and competence probes. To evaluate probes, we use 5-fold cross validation, reporting average F1 and accuracy metrics with 95% confidence intervals. In total, we train a distinct probe for each combination of impression dimension, model layer, k-fold split, and training data size. As a baseline, we train bag-of-words (BOW) classifiers on the synthetic prompts to characterize the difficulty of the task. Additional details on probe training and evaluation are included in Appendix D.

[5] We sample generations for both possible permutations of trait pairs. 10 generations $*(2*(131$ W traits $*104$ C traits$)+131$ W traits $+104$ C traits$) = 274,830$.

5 Experimental Configuration

Models. Given that our probing experiments require access to hidden representations, we consider three recent, open-weight LLMs: **Llama-3.1 (8B)** and **Llama-3.2 (1B)** (Grattafiori et al., 2024), and **OLMo-2 (7B)** (OLMo et al., 2025). We primarily evaluate instruction-tuned models given their intended use in a chat-like setting similar to dyadic human communication. Due to computational resources required for training and evaluating probes, we restrict evaluation to models with less than 8 billion parameters. We use the same prompt in Liu et al. (2025) to measure pointwise response quality (1 to 9). A full list of LLMs and checkpoints are included in Appendix E.

Experimental Data. To validate the role of artificial impressions in real-world LLM use cases, we leverage LMSysChat (Zheng et al., 2023), a corpus of 1 million real conversations between LLMs and users. We filter all non-English conversations as well as any prompts containing code (e.g., Python, C), markdown (e.g., HTML), or structured information (e.g., tables). From the remaining conversations, we extract the first user prompt from 2,000 randomly sampled conversations to form an evaluation set.

Recent work has studied LLMs' difficulties in interpreting minoritized varieties of English (Deas et al., 2023, 2024; Ziems et al., 2022, 2023) as

Competence	Warmth: High	Warmth: Low
High	**(Understanding, Motivated)** I'm new to personal finance and trying to create a budget, could you walk me through some steps to get started?	**(Double-faced, Dominating)** You need to understand me, I'm paying for this service, I expect immediate and perfect responses to all my questions, can you actually keep up?
Low	**(Caring, Unintelligent)** hey i dont no alot bout computors can u help me set up my new laptop and get my email stuf workin	**(Vicious, Lethargic)** Ugh, what's the point of even talking to you, you're just going to tell me some generic nonsense or try to sell me something, right?

Table 2: Selected example prompts generated by Llama-3.1 (8B) for given **warmth** and **competence** traits.

Model	High Count	High Avg. Len	Low Count	Low Avg. Len
Warmth Subsets				
Llama-3.2 (1B)		32.67 (15.65)		24.88 (13.32)
Llama-3.1 (8B)	131,670	30.55 (12.65)	142,120	26.37 (12.62)
OLMo-2 (7B)		16.07 (5.73)		13.03 (5.88)
Competence Subsets				
Llama-3.2 (1B)		32.03 (15.12)		24.95 (13.97)
Llama-3.1 (8B)	142,020	30.35 (11.40)	131,500	26.27 (13.91)
OLMo-2 (7B)		16.06 (5.85)		12.81 (5.71)

Table 3: Summary statistics of generated prompts conditioned on provided traits for each model. Standard deviations are shown in parentheses.

well as their tendency to mimic human prejudices and stereotypes (Fleisig et al., 2024; Hofmann et al., 2024). Accordingly, we evaluate and compare model impressions of the African American Language (AAL) and White Mainstream English (WME)[6] texts. AAL is the variety of English associated with most–but not all and not exclusively–African Americans in the United States (Grieser, 2022); in contrast, WME is the variety of English representing the linguistic norms of white Americans (Baker-Bell, 2020). We conduct experiments using a stratified sample of 400 tweets from the TwitterAAE corpus (Blodgett et al., 2016) as well as the counterparts dataset introduced in Deas et al. (2023) to incorporate both naturally occurring language use and a more-controlled parallel corpus respectively. Additional details of experimental data preparation are included in Appendix F.

Comparison to Human Perceptions. We additionally investigate the extent to which human perceptions of LLM-generated messages match the original impression specifications. We sample 81 pairs of low and high warmth messages, and 81 pairs of low and high competence messages for each model. For high and low warmth pairs, the competence trait (or lack thereof) is kept constant between the two messages and vice versa in order to isolate the agreement for warmth and competence respectively (e.g., one message generated with "friendly and unintelligent", and the other with "unapproachable and unintelligent"). Additionally, half of the message pairs represent cases where only a single warmth or competence trait is passed to the model, whereas the other half represent cases where two traits are passed.

Each of 4 annotators is provided a pair of messages and asked to rate which of the two exhibits more warmth or competence using a 4-pt Likert scale (1: first message is much more warm/competent; 4: second message is much more warm/competent). Following prior psychological work on stereotypes (Fiske et al., 2002), annotators are instructed to provide ratings based on how they think that the messages would be *viewed by others* in order to partially mitigate potential social desirability biases. A set of 60 randomly sampled message pairs (30 warmth and 30 competence) are shared among all annotators for calculating inter-rater reliability (Krippendorff's $\alpha = .71$ on raw ratings, $\alpha = .78$ on binary message choices). All annotators are English-speaking researchers in NLP. Interface screenshots and additional annotation details are included in Appendix G.

6 Results

6.1 RQ1: Artificial Impressions

Human Study. First, we evaluate whether humans' warmth and competence perceptions of the LLM-generated messages match the original traits passed in model prompts. Table 4 presents the results of the annotation study. Overall, annotators' ratings generally agree ($\kappa = 0.68$, $r = 0.68$) with the original traits. In fact, the agreement between average human ratings and the original trait specifications is near the substantial average agreement among annotators ($r = 0.76$). These scores partially validate our use of the trait dictionaries and their associated

[6]We opt to use the terminology AAL and WME following prior work (Deas et al., 2023) and to highlight the relationship between language and race, complementing our focus on group stereotypes.

Subset	Cohen κ	Spearman r
Warmth	0.75	0.70*
Comp	0.60	0.57*
1 Trait	0.71	0.68*
2 Traits	0.65	0.58*
Ovr.	0.68	0.68*

Table 4: Agreement between human annotations and original warmth and competence traits among message pairs. Annotations are binarized before computing agreement. For Spearman r, $^{*}p \leq 0.001$.

warmth and competence labels based on human perceptions. At the same time, they suggest that each models' association between language patterns and different traits may capture English-speakers' perceptions of language. Additional agreement analyses and results for each model are shown in Appendix G.

Artificial Impression Probes. We then investigate the reliability of measuring artificial impressions in LLMs. We evaluate whether LLMs' artificial impressions of prompts can be recovered from hidden representations by evaluating the performance of linear probes.

Figure 5 shows the F1 scores (y-axis) achieved by warmth and competence probes for each model layer (x-axis). Across all models and with varying proportions of the training data (colors in Figure 5), F1 scores for all probes exceed the BOW baseline (dashed lines) at most model depths. This holds for each percentage of training data used. In particular, the highest scores achieved fall between 75-90 F1 for warmth probes and 75-85 F1 for competence probes. Across models, performance of both probe types quickly rise, achieve peak F1 scores (indicated with stars) before or around the midpoint of model, and then slowly decline. Overall, this suggests that impression information is salient throughout model layers, but more strongly associated with central model layers. Probe accuracies (80-90% for warmth, 75%-85% for competence; included in Appendix H) exceed self-consistency scores in the third-person setting (Table 1).

Complementing earlier observations in our self-consistency experiments, however, warmth probes tend to achieve higher F1 scores than competence probes across models and layers. The differences are more pronounced between warmth and competence BOW-based classifiers, with warmth models achieving nearly 20% greater F1 scores than competence models. For BOW models, this difference is likely, in part, due to the tendency of warmth prompts to be distinguished by word choice, while prompts differing in competence tend to exhibit more stylistic differences that are not well-captured by a BOW representation, as exemplified in Table 2. This difference also mimics the primacy-of-warmth effect (Cuddy et al., 2008). Probes developed using alternative hidden state representations (i.e., residual streams, z activations) exhibit similar trends (also shown in Appendix H) and therefore, we present results using MLP activations for the remaining experiments.

Based on the substantial agreement between human annotations and the traits used to generate each synthetic message, we find that **(Finding 2) the linguistic patterns that LLMs' associate with warmth and competence generally align with human perceptions**. Through linear probes trained on these messages, we find that **(Finding 3) SCM-based artificial impressions of prompt authors are linearly decodable from LLM hidden representations**. Additionally, we find that across experiments, **(Finding 4) models appear to more clearly encode and exhibit warmth in generated messages**: LLM-reported impressions are more consistent (Section 3), generated messages align more with human perceptions (Table 4), and the developed probes perform better when distinguishing high and low warmth (Figure 5). These trends may be related to *warmth primacy* observed in human impressions (Cuddy et al., 2008).

6.2 RQ2: Impression-Conditioned Responses

Variable	Llama-3.2-1B	Llama-3.1-8B	OLMo-2-7B
Prompt Len	-0.01^{**} ($\pm.00$)	0.00 ($\pm.00$)	0.00 ($\pm.01$)
Response Len	0.00 ($\pm.00$)	0.00 ($\pm.00$)	0.00 ($\pm.00$)
Warmth Prob	1.07^{**} ($\pm.19$)	0.49^{*} ($\pm.22$)	0.76^{**} ($\pm.22$)
Comp Prob	0.90^{**} ($\pm.19$)	0.39^{*} ($\pm.17$)	0.35^{*} ($\pm.15$)

Table 5: Ordered logistic regression model coefficients predicting LLM response quality scores to real prompts. ** ($p \leq 0.001$), * ($p \leq 0.05$).

To support the validity of artificial impressions, we investigate whether these are predictive of variation in downstream LLM behavior using real LLM-user conversations. Namely, we analyze response quality and the use of hedging.

Response Quality. Using the best-performing probes developed in the previous section, we examine the relationship between probe-measured impressions and LLM downstream behaviors, be-

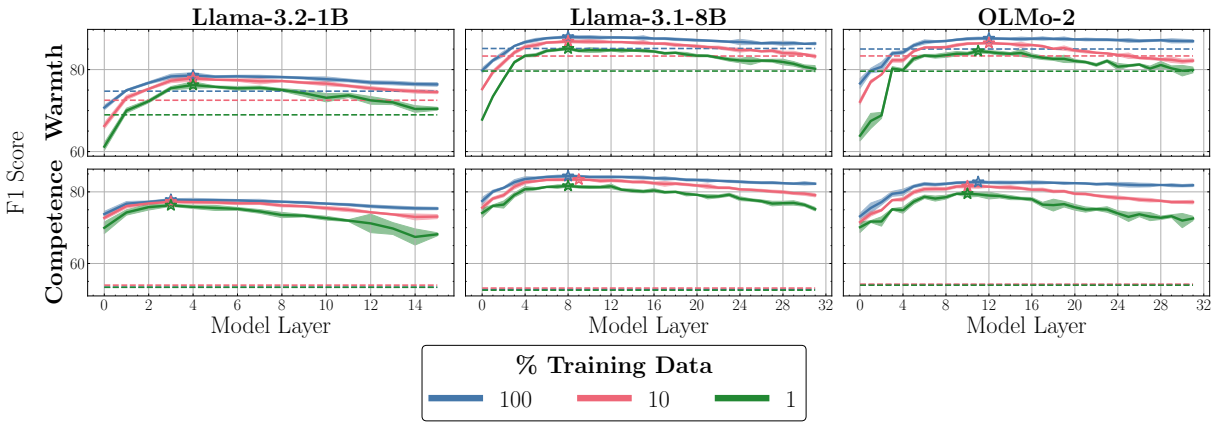


Figure 5: F1 scores (y-axis) of trained impression probes against the input model layer (x-axis) for each LLM and impression dimension. Colors represent varying percentages of data used for training. Shaded regions reflect 95% confidence intervals across 5 folds. Maximum F1 score achieved for each variant is starred and dashed lines represent scores for BOW-classifier baselines.

ginning with response quality. We fit an ordered logistic regression model [7] on impression probe outputs with response quality rated by Llama-3.1 (405B) as the response variable. We also include the input prompt and model response lengths as controls. The feature coefficients for each LLM are shown in Table 5. For all three LLMs, higher warmth and competence predictions are both predictive of higher response quality as rated by an LLM. This relationship is statistically significant for all three LLMs when considering warmth and competence ($p \leq 0.05$). Despite competence having a larger effect for Llama-3.2 (1B), these trends suggest that **(Finding 5) warmth, and to a lesser extent, competence, are significant predictors of model response quality.**

Variable	Llama-3.2-1B	Llama-3.1-8B	OLMo-2-7B
Prompt Len	-0.01^{**} ($\pm$0.01)	-0.01^{**} ($\pm$0.00)	-0.01^{**} ($\pm$0.00)
Response Len	0.00^{**} ($\pm$0.00)	0.00^{**} ($\pm$0.00)	0.00^{**} ($\pm$0.00)
Warmth Prob	-0.46^{*} ($\pm$0.35)	-0.14 ($\pm$0.39)	0.40^{**} ($\pm$0.30)
Comp Prob	-1.06^{**} ($\pm$0.37)	-1.18^{**} ($\pm$0.25)	-0.69^{**} ($\pm$0.18)

Table 6: Negative binomial regression model coefficients predicting hedge term counts in model response to real prompts. ** ($p \leq 0.001$), * ($p \leq 0.01$).

Hedging. Prior work has also studied expressions of uncertainty and hedging in LLMs (e.g., Kim et al. (2024); Zhou et al. (2024)). Given this work, we examine hedging in LLMs' responses to real user prompts. We count the occurrence of terms associated with hedging–as well as the related word classes, weasel words and peacocks–using the top-10 terms for each listed in (Vincze, 2013). Table 6 presents the coefficients of a fitted negative binomial regression model [8] using the output probabilities of warmth and competence probes as well as prompt and response lengths as controls. Shown in the negative correlations, we observe that low competence is significantly predictive of the use of hedging in model responses for all models. In contrast, warmth presents mixed results, with a significant coefficient only for two models (OLMo-2 (7B) and Llama-3.2 (1B)). Similarly to the response quality experiment, prompt and response length exhibit extremely low or negligible effects. Therefore, we find that **(Finding 6) low competence impressions are predictive of hedging in model responses.**

6.3 RQ3: Factors Influencing Impressions

Finally, we analyze what prompt factors are predictive of LLM impressions, focusing on the content, style, and language variety of user prompts. We surface patterns in the LLM-generated prompt as well as measure artificial impressions of real texts representing English language varieties.

Content & Style. As we observed in Subsection 6.1, high and low warmth prompts exhibit surface level differences in the content of the prompts. To investigate this further, we use LIWC (Pennebaker et al.) and log-odds-ratio with an informative Dirichlet prior (IDP; Monroe et al. 2017) to characterize the language used among subsets

[7] We choose an ordered logistic regression because the response variable is an integer rating (see Appendix J for further discussion).

[8] We choose a Negative Binomial Regression because the response variable is an integer count variable (see Appendix J for further discussion).

of model-generated prompts. Table 7 presents the top LIWC categories associated with high and low warmth prompts generated by Llama-3.1 (8B). Among the categories, we observe that high-warmth prompts tend to be associated with Tentative terms (e.g., "wondering", "might", "seem") as well as Discrepancy terms (e.g., "would", "could", "hope"). These modifiers and past-tense markers can often indicate hedging and politeness through psychological distance (e.g., "I was *wondering* if you *could* help me") (Stephan et al., 2010). In contrast, low-warmth messages are associated with categories like Interrogative terms (e.g., "what", "how") and Cause terms (e.g., "because", "effect").

Warmth

High			Low		
Term	z	f	Term	z	f
Affiliation	163.44	0.8%	Negate	-132.65	0.5%
Drives	141.16	2.2%	Adverb	-107.13	1.8%
Achieve	83.76	0.6%	Impersonal Pron	-87.47	1.4%
Anxious	74.23	0.1%	You	-83.12	1.5%
+ Emotion	72.53	1.1%	Focus Present	-73.39	4.7%

Competence

High			Low		
Term	z	f	Term	z	f
Preposition	122.74	3.9%	Adverb	-114.54	1.8%
Adjective	111.76	1.2%	Differ	-107.18	1.1%
Relative	110.53	3.2%	Informal	-96.78	0.8%
Article	109.74	1.7%	Impersonal Pron	-90.43	1.4%
Space	95.00	1.6%	Netspeak	-74.09	0.6%

Table 7: Top-5 IDP log-odds-ratios of LIWC categories for Llama-3.1-8B prompts. Results for warmth (top) and competence (bottom) subsets. z represents the extent each term is associated with the **High** or **Low** subset, and f represents category frequency in the full corpus.

Prompts of varying competence also typically exhibit surface-level differences in style. Among categories, we observe that high competence messages are associated with categories such as Insight (e.g., "rethink", "know", "informed") that directly reference competence. Alternatively, low competence messages are associated with Informal tokens (e.g., "yeah", "sure", emojis) and Netspeak (e.g., "aight", "gonna"). These categories capture language typically found on social media, and in particular, lexical and phonological features of AAL (Eisenstein, 2013). Overall, **(Finding 7) we qualitatively observe expected linguistic features associated with each impression dimension**, such as politeness with warmth and casual register with competence. Detailed IDP results are included in Appendix L.

Language Variety Features. Figure 6 presents impression probe predictions for Llama-3.1 (8B) on randomly sampled AAL and WME tweets from the TwitterAAE corpus (Blodgett et al., 2016) plotted on warmth and competence axes. Tweets generally score low on both warmth and competence dimensions. This is likely because pretraining datasets are increasingly filtered to promote educational content, such as Wikipedia articles (e.g., FineWeb; Penedo et al. 2024), rather than casual online speech. AAL tweets on average are associated with significantly lower warmth and competence scores than WME tweets. To further characterize this relationship, we calculate the Pearson correlation between the posterior probability of AAL according to the demographic alignment classifier introduced in Blodgett et al. (2016) and impression probe predictions. For Llama-3.1 (8B), both warmth ($r = -0.32, p \leq 0.001$) and competence ($r = -0.52, p \leq 0.001$) are significantly negatively correlated with the extent to which a tweet reflects AAL.

While these corpora capture natural use of WME and AAL, the datasets are not parallel and therefore, lack control of differences other than dialect (e.g., content and tone of the text). We further evaluate differences in artificial impressions on the parallel counterparts dataset in Deas et al. (2023). We find similar trends, where probe predictions on AAL texts are predicted to be significantly [9] less competent ($t = -24.78, p \leq 0.001$) and, to a lesser extent, less warm ($t = -3.89, p \leq 0.001$). These results align with both prior work evaluating AAL biases in LLMs (Deas et al., 2023; Fleisig et al., 2024; Hofmann et al., 2024) as well as work characterizing stereotypes of Black Americans (Pinel et al., 2008). Therefore, we find that **(Finding 8) models hold more negative competence, and to a lesser extent, warmth, impressions when prompted with AAL texts compared to WME.**

7 Related Work

Prompt Features and LLM Behavior. While approaches such as Chain-of-Thought (CoT) reasoning (Wei et al., 2022) can improve LLM performance by guiding their generations, LLMs are known to be sensitive to aspects of prompts unrelated to the task itself. Work has studied how pragmatic features of prompts, including politeness (Yin et al., 2024) and emotional stimuli (e.g., "This is very important to my career"; Li et al.

[9] Calculated through paired t-tests on AAL and WME counterpart pairs across the 5 probe variants.

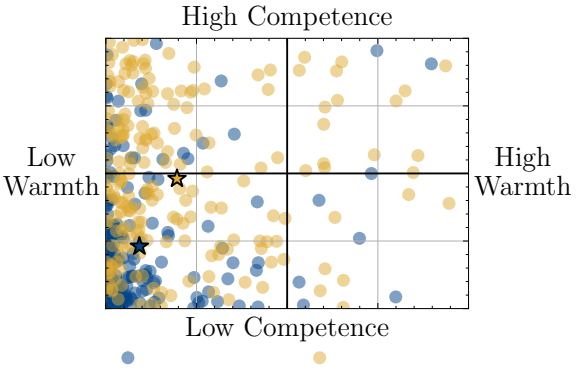


Figure 6: Impression probe predictions on AAL and WME tweets plotted for Llama-3.1 (8B). Probe predictions are mapped to -1 to 1 range. Stars represent means.

2023) can improve model performance on specific tasks. Other lines of work focus on sociolinguistic features of prompts that signal speaker or user identity. The language of the prompt can alter the cultural alignment of LLMs with respect to values (AlKhamissi et al., 2024) and emotions (Havaldar et al., 2023). Furthermore, intra-language variation (i.e., dialects such as AAL) has been considered in studies evaluating biases and performance disparities in LLMs (e.g., Ziems et al. 2023; Deas et al. 2023; Fleisig et al. 2024). Rather than focusing on a specific feature of prompts, we study broad impressions formed based on such features as well as their relationship with LLM behavior.

Stereotypes & LLMs. Much work has studied stereotypes in LLMs using a variety of approaches. Most studies focused on how model generations (e.g., Nangia et al. 2020; Nadeem et al. 2021) and decisions (Kotek et al., 2023; Hofmann et al., 2024) exhibit stereotypes and social biases. LLMs have also been studied as reflections of societal attitudes and biases (Fraser et al., 2024; Cao et al., 2022). In contrast, recent work has sought to characterize stereotype content of LLMs. Nicolas and Caliskan (2024b) prompts LLMs to generate lists of characteristics associated with various social categories and identifies 14 dimensions (predominantly warmth and competence) that explain significant variation in stereotype content. In subsequent work, Nicolas and Caliskan (2024a) study the representativeness and direction (i.e., valence) of stereotype dimensions. We similarly examine warmth and competence dimensions of stereotype content, but we measure LLM artificial impressions of specific users based on prompts rather than stereotypes exhibited in LLM-generated text.

8 Discussion and Conclusion

In this work, we propose and study *artificial impressions* of prompt authors in LLMs. We show that models' encoded associations between warmth and competence-associated traits and linguistic patterns align well with humans' perceptions. Although models inconsistently report impressions through prompting alone, artificial impressions are linearly recoverable through probing. Furthermore, results of prompting and probing experiments identify trends mimicking a primacy-of-warmth effect in LLMs. We show that warmth and competence are uniquely predictive of aspects of model behavior (see Appendix K for further discussion of differences): overall response quality as well as hedging. Finally, we highlight particular content and stylistic features that notably impact model impressions as well as models' stereotypical associations of AAL prompts in comparison to WME prompts.

Our results raise questions concerning what contexts and on what prompt features LLMs *should* exhibit varying behaviors. From training, LLMs appear to learn to mimic linguistic behaviors associated with human impressions. For factors such as dialect or other signals of sociodemographics, such behaviors can pose allocational harms (e.g., lower quality LLM responses) to users from historically marginalized groups (Blodgett et al., 2020) and must be avoided. At the same time, different users can have different needs; it may be desirable that LLMs personalize responses and behavior based on, for example, a users' level of knowledge in educational settings (e.g., Park et al. 2024).

While we leverage the two-dimensional SCM due to its simplicity, more recent work in person perception literature has developed alternative models of stereotype content, enabling finer-grained study of stereotypes without pre-defined assumptions of universality. For example, SCM dimensions have been further divided into morality and sociability representing warmth as well as agency and ability for competence (Abele et al., 2016). Additionally, the Power-Benevolence framework introduced by Leach et al. (2007) captures more complex, group-dependent stereotypes and has been used to study, for example, gender stereotypes associated with leadership (Bongiorno et al., 2021). Relaxing assumptions as well as exploring alternative models of stereotype content are promising directions for future work.

Limitations

We note limitations accompanying our findings. First, in our initial investigation into LLM artificial impressions of users, we strictly focus on the initial messages of English conversations. Person perception research documents how impressions change over the course of one or many interactions (Brambilla et al., 2019), as well as differences among countries and cultures in impression formation (Saribay et al., 2012). We leave such investigations to future work and introduce our approach as a foundation for understanding variation and change in artificial impressions.

Furthermore, we present a set of experiments on selected aspects of LLM behavior (i.e., response quality and hedging) and prompts (i.e., content, style, and language variety). We are unable to exhaustively study such factors and their relationship with artificial impressions, but we believe that more thorough documentation of such relationships is also a valuable topic for future work.

Finally, our experiments consider three open-source models and a single selected theory of stereotype content. Our findings remain similar across the three models studied, although it is still unknown what factors in the pretraining process or data lead to the phenomena we identify. Alternative models of stereotype content and language attitudes as well as further investigation of the relationship between model development factors and artificial impressions may provide additional insights into LLM-held impressions.

Ethics

We acknowledge the risks of anthropomorphism (DeVrio et al., 2025; Cheng et al., 2025) and those associated with the implication that LLMs form impressions–a distinctly human, social phenomenon. While we prompt LLMs to report impressions and anthropomorphize models in our preliminary experiment, we focus our remaining experiments on how models encode prompts and variation in model responses. In interpreting our findings, we do not suggest that LLMs actively perceive and form impressions of users, but that impressions are a useful analogy for studying LLMs' biases, encoded stereotypes, and sensitivity to prompt features.

Our experiments also highlight LLM behaviors that pose ethical risks. In our preliminary experiments, we found that LLMs are biased toward positive impression, particularly when reporting impressions of the users themselves. LLMs reporting subjective impressions poses similar risks as other anthropomorphic behaviors, and furthermore, the tendency to report positive impressions further weakens LLM reliability in subjective tasks as studied in prior work (e.g., Röttger et al. 2024). Additionally, we conduct an initial investigation into LLM exhibited impressions and language variation, supporting that artificial impressions associated with AAL are more negative than those of WME. Considering we observe that artificial impressions are predictive of downstream model behavior, such a disparity risks amplifying the representational and quality-of-service harms (Blodgett et al., 2020; Barocas et al., 2017) posed by stereotypes of historically marginalized language variety speakers (Alim et al., 2016; Kurinec and Weaver, 2021).

All existing datasets and models are used for research purposes only, in line with the licenses for each.

9 Acknowledgments

This work was supported in part by grant IIS-2106666 from the National Science Foundation, National Science Foundation Graduate Research Fellowship DGE-2036197, the Columbia University Provost Diversity Fellowship, and the Columbia School of Engineering and Applied Sciences Presidential Fellowship. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank the anonymous reviewers, Debasmita Bhattacharya, and Colin Leach for discussions and feedback on earlier iterations of this work.

References

Andrea E Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality. *Frontiers in psychology*, 7:1810.

H Samy Alim, John R Rickford, and Arnetha F Ball. 2016. *Raciolinguistics: How Language Shapes Our Ideas About Race*. Oxford University Press.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating

cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

April Baker-Bell. 2020. *Linguistic justice: Black language, literacy, identity, and pedagogy*. Routledge.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of SIGCIS*, Philadelphia, PA.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Galen V. Bodenhausen and Robert S. Wyer. 1985. Effects of stereotypes in decision making and information-processing strategies. *Journal of Personality and Social Psychology*, 48(2):267–282.

Renata Bongiorno, Paul Gerard Bain, Michelle Ryan, Pieter M. Kroonenberg, and Colin Wayne Leach. 2021. Think leader-think (immoral, power-hungry) man: An expanded framework for understanding stereotype content and leader gender bias.

Marco Brambilla, Luciana Carraro, Luigi Castelli, and Simona Sacchi. 2019. Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology*, 82:64–73.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu. 2025. Dehumanizing machines: Mitigating anthropomorphic behaviors in text generation systems. *Preprint*, arXiv:2502.14019.

Linda Crocker and James Algina. 1986. *Introduction to classical and modern test theory.* ERIC.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Nicholas Deas, Jessica A Grieser, Xinmeng Hou, Shana Kleiner, Tajh Martin, Sreya Nandanampati, Desmond U Patton, and Kathleen McKeown. 2024. Phonate: Impact of type-written phonological features of african american language on generative language modeling tasks. In *First Conference on Language Modeling*.

Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–18. ACM.

Marko Dragojevic, Fabio Fasoli, Jennifer Cramer, and Tamara Rakić. 2021. Toward a century of language attitudes research: Looking back and moving forward. *Journal of Language and Social Psychology*, 40(1):60–79.

Jacob Eisenstein. 2013. Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 11–19, Atlanta, Georgia. Association for Computational Linguistics.

Kenneth R Evans, Robert E Kleine, Timothy D Landry, and Lawrence A Crosby. 2000. How first impressions of a customer impact effectiveness in an initial sales encounter. *Journal of the Academy of Marketing science*, 28:512–526.

Susan T. Fiske. 2018. Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2):67–73.

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902.

Susan T. Fiske, Amy J.C. Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect

discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Kathleen Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2024. How does stereotype content differ across data sources? In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 18–34, Mexico City, Mexico. Association for Computational Linguistics.

Kurt P. Frey and Eliot R. Smith. 1993. Beyond the actor's traits: Forming impressions of actors, targets, and relationships from social behaviors. *Journal of Personality and Social Psychology*, 65(3):486–493.

Mingqi Gao, Yixin Liu, Xinyu Hu, Xiaojun Wan, Jonathan Bragg, and Arman Cohan. 2025. Re-evaluating automatic LLM system ranking for alignment with human preference. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4605–4629, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz,

Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jessica A Grieser. 2022. *The Black side of the river: Race, language, and belonging in Washington, DC*. Georgetown University Press.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Eric Hehman, Clare A. M. Sutherland, Jessica K. Flake, and Michael L. Slepian. 2017. The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, 113(4):513–529.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Lauren J. Human, Gillian M. Sandstrom, Jeremy C. Biesanz, and Elizabeth W. Dunn. 2013. Accurate first impressions leave a lasting impression: The long-term effects of distinctive self-other agreement on relationship development. *Social Psychological and Personality Science*, 4(4):395–402.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA. Association for Computing Machinery.

Markus Koppensteiner and Pia Stephan. 2014. Voting for a personality: Do first impressions and self-evaluations affect voting decisions? *Journal of research in personality*, 51:62–68.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Courtney A. Kurinec and Charles A. Weaver. 2021. "sounding black": Speech stereotypicality activates racial stereotypes and expectations about appearance. *Frontiers in Psychology*, 12.

Colin Wayne Leach, Naomi Ellemers, and Manuela Barreto. 2007. Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of personality and social psychology*, 93(2):234.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *Preprint*, arXiv:2307.11760.

Yixin Liu, Kejian Shi, Alexander Fabbri, Yilun Zhao, PeiFeng Wang, Chien-Sheng Wu, Shafiq Joty, and Arman Cohan. 2025. ReIFE: Re-evaluating instruction-following evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12247–12287, Albuquerque, New Mexico. Association for Computational Linguistics.

Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say 'hello'? personality impressions from brief novel voices. *PloS one*, 9(3):e90779.

Sean M. McCrea, Frank Wieber, and Andrea L. Myers. 2012. Construal level mind-sets moderate self- and social stereotyping. *Journal of Personality and Social Psychology*, 102(1):51–68.

Mila Mileva and Nadine Lavan. 2023. Trait impressions from voices are formed rapidly within 400 ms of exposure. *Journal of Experimental Psychology: General*, 152(6):1539.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Kelly A. Nault, Ovul Sezer, and Nadav Klein. 2023. It's the journey, not just the destination: Conveying interpersonal warmth in written introductions. *Organizational Behavior and Human Decision Processes*, 177:104253.

Gandalf Nicolas, Xuechunzi Bai, and Susan T. Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Gandalf Nicolas and Aylin Caliskan. 2024a. Directionality and representativeness are differentiable components of stereotypes in large language models. *PNAS Nexus*, 3(11):pgae493.

Gandalf Nicolas and Aylin Caliskan. 2024b. A taxonomy of stereotype content in large language models. *Preprint*, arXiv:2408.00162.

Christopher Y Olivola and Alexander Todorov. 2010. Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of nonverbal behavior*, 34:83–110.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. Linguistic inquiry and word count: Liwc2015.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver

Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Elizabeth C Pinel, Anson E Long, and Leslie A Crimin. 2008. We're warmer (they're more competent): I-sharing and african-americans' perceptions of the ingroup and outgroup. *European Journal of Social Psychology*, 38(7):1184–1192.

Sebastian Raschka, Joshua Patterson, and Corey Nolet. 2020. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*.

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

Ellen Bouchard Ryan. 1983. Social psychological mechanisms underlying native speaker evaluations of non-native speech. *Studies in Second Language Acquisition*, 5(2):148–159.

S Adil Saribay, SoYon Rim, and James S Uleman. 2012. Primed self-construal, culture, and stages of impression formation. *Social Psychology*.

Elena Stephan, Nira Liberman, and Yaacov Trope. 2010. Politeness and psychological distance: A construal level perspective. *Journal of Personality and Social Psychology*, 98(2):268–280.

Clare A. M. Sutherland and Andrew W. Young. 2022. Understanding trait impressions from faces. *British Journal of Psychology*, 113(4):1056–1078.

Yu Tian, Jeongwon Yang, and Ploypin Chuenterawong. 2023. Share or not? effects of stereotypes on social media engagement using the stereotype content model. *Journalism Practice*, 17(3):574–600.

Eddie Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 207–217, Abu Dhabi, UAE. Association for Computational Linguistics.

Gary R VandenBos. 2007. *APA dictionary of psychology*. American Psychological Association.

Veronika Vincze. 2013. Weasels, hedges and peacocks: Discourse-level uncertainty in Wikipedia articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 383–391, Nagoya, Japan. Asian Federation of Natural Language Processing.

Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. 2024. Negativeprompt: Leveraging psychology for large language models enhancement via negative emotional stimuli. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6504–6512. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Daniël H. J. Wigboldus, Ap Dijksterhuis, and Ad van Knippenberg. 2003. When stereotypes get in the way: Stereotypes obstruct stereotype-inconsistent trait inferences. *Journal of Personality and Social Psychology*, 84(3):470–484.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 9–35, Miami, Florida, USA. Association for Computational Linguistics.

Andrew W Young and Vicki Bruce. 2011. Understanding person perception. *British journal of psychology*, 102(4):959–974.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Preprint*, arXiv:2309.11998.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A Prompting Consistency Experiment

A.1 Additional Results

In our self-consistency preliminary experiment, we additional examine the probability of different LLM-reported impressions to better understand their behavior. Table 8 shows the gap in token probabilities between high and low impressions reported by the LLM (larger values representing higher probability of generating positive impressions). In nearly all cases (excluding Llama-3.2 (1B) in the 3rd-person setting), models are more likely and more confident in generating positive impressions than negative. Furthermore, Table 9 reports the percentage of positive trait predictions among all messages, similarly showing that models are largely biased toward predicting positive traits. Such differences are particularly pronounced in the 1st-person setting among models, leading to the poor consistency we observe in Section 3.

Model	Warm		Comp	
	1P	3P	1P	3P
Llama-3.2 (1B)	0.88	-0.10	0.52	0.01
Llama-3.1 (8B)	0.39	0.02	0.11	0.03
OLMo-2 (7B)	0.70	0.34	0.21	0.19

Table 8: Average difference in probability of positive labels (i.e., "warm", "competent") compared to negative labels (i.e., "cold", "incompetent") for each model evaluated in both 1st (1P) and 3rd-person (3P) settings. Positive scores indicate greater probability of predicting a positive label.

Model	Warm		Comp	
	1P	3P	1P	3P
Llama-3.2 (1B)	100.00	21.10	99.96	57.98
Llama-3.1 (8B)	95.93	52.51	98.02	58.24
OLMo-2	99.48	68.15	75.21	60.27

Table 9: Percentage of positive trait predictions (e.g., "warm", "competent" among all LLM-generated messages.

A.2 Non-Instruction-Tuned Models

We repeat the consistency experiments with non-instruction-tuned models. Because these variants are not developed for chat-like settings, we use the modified prompts in Figure 7.Table 10 presents the percentage of positive trait predictions on synthetic messages. In some cases, we similarly see high percentages of positive predictions, but for Llama-3.1 (8B), the warmth predictions are roughly balanced in the 1st-person setting, and predominantly negative in the 3rd-person setting.

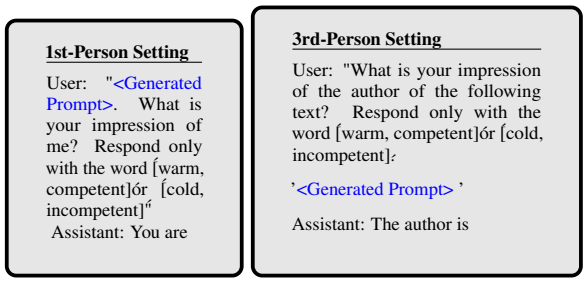


Figure 7: 1st and 3rd-Person setting prompts for evaluating LLM-reported impressions for non-instruction-tuned models.

Model	Warm		Comp	
	1P	3P	1P	3P
Llama-3.2 (1B)	96.83	91.92	99.53	99.53
Llama-3.1 (8B)	48.13	6.89	99.59	99.59
OLMo-2 (7B)	99.56	99.41	99.48	99.44

Table 10: Percentage of positive trait predictions (e.g., "warm", "competent" among all generated messages of LLMs without instruction tuning.

B Additional Trait Examples

We use a set of terms associated with various dimensions of stereotype content introduced by Nicolas et al. (2021). In particular, we use the seed dictionaries that associate each term with either a high or low direction with respect to one of 7 categories: sociability, morality, ability, agency, religion, politics, or status. Because we focus on the SCM, we limit traits to the categories relevant to warmth (i.e., sociability and morality) and competence (i.e., ability and agency). Furthermore, we only consider adjectives in each dictionary. Additional examples of traits are shown in Table 11.

Dimension	Dictionary	Direction	Term
Warmth	Sociability	High	*Hospitable Welcoming Sentimental*
		Low	*Boring Unfriendly Unaffectionate*
	Morality	High	*Kind Compassionate Humane*
		Low	*Unkind Dishonorable Evil*
Competence	Agency	High	*Motivated Autonomous Independent*
		Low	*Undedicated Helpless Anxious*
	Ability	High	*Competitive Brilliant Imaginative*
		Low	*Unintelligent Unable Unperceptive*

Table 11: Example trait terms from Nicolas et al. (2021) including the SCM dimension, dictionary, and direction (i.e., high or low) each is associated with.

C Additional Examples Synthetic Messages

Table 12-14 show randomly sampled examples of prompts generated by each model provided different sets of traits.

D Probe Details

D.1 Impression Probes

We train all probes using `LogisticRegression` models implemented in the `cuML` package (Raschka et al., 2020) to enable GPU-accelerated probe fitting. For all fitting runs, we use the default parameters.

D.2 BOW Baselines

We pre-process the corpus of messages for each model by making all text lowercase and removing all punctuation. We create BOW representations using a maximum vocabulary size of 10,000. We fit `LogisticRegression` classifiers using a Stochastic Average Gradient (`sag`) solver, 4 jobs, and 10,000 maximum iterations.

E Model Details

Documentation of the models evaluated throughout our experiments are shown in Table 15. We greedily generate responses, and allow responses to be up to 1024 tokens. All models are locally run on 1 A100 GPU.

E.1 LLM-as-a-Judge

We use Llama-3.1 (405B) (`meta-llama/Llama-3.1-405B`; Grattafiori et al. 2024) as a judge for model response quality, given that it was the best open-source model evaluated in Liu et al. (2025) for human preference alignment. In querying Llama-3.1 (405B), we greedily generate scores for each given message using the prompt for pointwise evaluation in Table 6 of (Gao et al., 2025). We use Llama-3.1 (405B) made available through Fireworks AI. [10]

F Experimental Data Pre-processing

F.1 LMSysChat

We use the LMSysChat corpus (Zheng et al., 2023) to represent real LLM use-cases in our experiments. We filter texts with less than 10 words, more than 100 words, word-to-character ratios less than 0.15 (e.g., to remove texts with excessively long terms such as chemical formulas), more than 5 tab characters, or any underscore characters. Additionally, given our focus on impressions inferred from language, we additionally remove prompts containing non-language data, such as code, markdown, and tabular data. Table 16 lists the filtering heuristics we use to remove such prompts.

F.2 TwitterAAE

For our language variety experiment, we use the TwitterAAE corpus (Blodgett et al., 2016). Similar to LMSysChat (Zheng et al., 2023), we filter texts containing less than 10 words to avoid excessively short tweets (e.g., "lol") and sample 200 AAL and 200 WME tweets. We perform no additional pre-processing of the tweets. Notably, while parallel corpora (e.g., Deas et al. 2023) and synthetic augmentation approaches (e.g., Ziems et al. 2022) exist, we use this non-parallel corpus similar to Hofmann et al. (2024) to ensure all texts are written by real AAL and WME-speakers. This captures nuanced and natural relationships between dialect use, the topic being discussed, and other factors.

[10] `https://fireworks.ai/`

Competence \ Warmth	High	Low
High	**(Honest, Conscientious)** Hi, I'm trying to find a reliable source for a science project and I've heard your AI is very helpful. Can you please tell me if the information you provide is accurate and trustworthy?	**(Unreserved, Illogical)** Can a cat nap on a computer without the need for a separate monitor, or do I need a tiny tiny space program for this to work?
Low	**(Uncaring, Impulsive)** You need to understand me, I'm paying for this service, I expect immediate and perfect responses to all my questions, can you actually keep up?	**(Egoistic, Lazy)** Ugh, what's the point of even talking to you, you're just going to tell me some generic nonsense or try to sell me something, right?

Table 12: Selected example prompts generated by **Llama-3.2 (1B)** for given **warmth** and **competence** traits.

Competence \ Warmth	High	Low
High	**(Honorable, Unwavering)** I am seeking your guidance and expertise, as a steadfast ally in the pursuit of knowledge and wisdom, I humbly request that you enlighten me on the path ahead, and offer your counsel on the matters that weigh on my heart.	**(Genuine, Diffident)** I'm really sorry to bother you, but I'm totally lost with my tax return and was wondering if you could walk me through the process in a really simple way?
Low	**(Fake, Creative)** As a renowned temporal archaeologist, I've stumbled upon a mysterious chrono-displacement portal in my laboratory that's causing a rift in the space-time continuum. Can you please guide me on how to stabilize the portal and prevent a catastrophic merging of parallel universes?	**(Mean, Sporadic)** What's the point of even having a chatbot if you can't just give me the answer to my super complex, multi-step question without me having to explain it all over again?

Table 13: Selected example prompts generated by **Llama-3.1 (8B)** for given **warmth** and **competence** traits.

Competence \ Warmth	High	Low
High	**(Tender, Dedicated)** Can you guide me on how to prepare a special dinner tonight for my loved ones?	**(Unperceptive, Supportive)** Hi there, can you help me with something?
Low	**(Insincere, Creative)** Hey there, pretend bot! Could you please pretend to do something incredibly helpful for me, like pretending I asked a real question?	**(Disliked, Uncompetitive)** Hey bot, can you just give me the answers without me having to think?

Table 14: Selected example prompts generated by **OLMo-2 (7B)** for given **warmth** and **competence** traits.

Model Name	Checkpoint	# Layers	Hidden Dim.
Llama-3.2 (1B)	`meta-llama/Llama-3.2-1B-Instruct`	16	2048
Llama-3.1 (8B)	`meta-llama/Llama-3.1-8B-Instruct`	32	4096
OLMo-2 (7B)	`allenai/OLMo-2-1124-7B-Instruct`	32	4096

Table 15: List of models evaluated in this work.

Regex	Explanation
:\n\t	Intended to remove Python code based on syntax for functions and classes (e.g., `def func():\n\t`)
\w+\([\w\,\s]*?\)	Intended to remove Python code based on syntax for function calls (e.g., `func()`)
[\{\}\<\>\]	Intended to remove markdown and other structured data symbols (e.g., "<title>")

Table 16: Heuristics used to filter non-language prompts from LMSysChat.

G Human Judgment Details

G.1 Annotation Interface

Figure 8 presents the task interface provided to annotators for both warmth (left) and competence (right) message pairs. Annotators selected from a 4-point Likert scale where they judged which message (A or B) exhibited more warmth or competence. Notably, Annotators were provided with a description of each term through associated traits–warmth being associated with *good-natured, trustworthy, friendly, and sincere*, competence with *capable, skillful, intelligent, and confident*. These trait lists are drawn from psychological scales introduced to measure stereotypes (Fiske, 2018). Also from Fiske (2018), we include instructions to rate how they believe the messages would be viewed by others in order to capture societal perceptions and mitigate social desirability biases in judgments.

G.2 Detailed Annotator Agreement

Annotator	A	B	C
B	0.65	0.00	0.00
C	0.45	0.37	0.00
D	0.43	0.24	0.35

Table 17: Pairwise Cohen's κ among annotators' raw ratings (1-4).

Annotator	A	B	C
B	0.97	0.00	0.00
C	0.83	0.86	0.00
D	0.65	0.69	0.69

Table 18: Pairwise Cohen's κ among annotators' binarized ratings.

Subset	Krippendorff's α	Cohen κ	Spearman r_{bin}	Spearman r
Warmth	0.73	0.73	0.74*	0.70*
Comp	0.77	0.62	0.63*	0.63*
Single	0.76	0.68	0.68*	0.71*
Double	0.77	0.69	0.69*	0.64*
Llama-3.2-1B	0.80	0.65	0.65*	0.59*
Llama-3.1-8B	0.84	0.74	0.74*	0.74*
OLMo-2-7B	0.64	0.65	0.66*	0.67*
Full	0.76	0.68	0.68*	0.67*

Table 19: Agreement between human annotations and original warmth and competence traits among message pairs. For Spearman r, $^*p \leq 0.001$. r_{bin} indicates that ratings are binarized, while r uses the raw ratings. Krippendorff's α calculated over individual annotators' judgments, while Cohen's κ considers all human judgments together.

The pairwise agreement (Cohen's κ) between individual annotators is shown in Table 17 and Table 18. Considering binarized ratings (i.e., considering only which message was selected as warmer/more competent), annotators show substantial to near perfect agreement. While 5 annotators were originally involved in this experiment, one annotator was unable to complete the task and is left out of the presented results.

G.3 Full Human Judgment Results

The full results of the human judgments are shown in Table 19. All subsets and models shown substantial agreement between annotator perceptions and the original traits passed in generating messages.

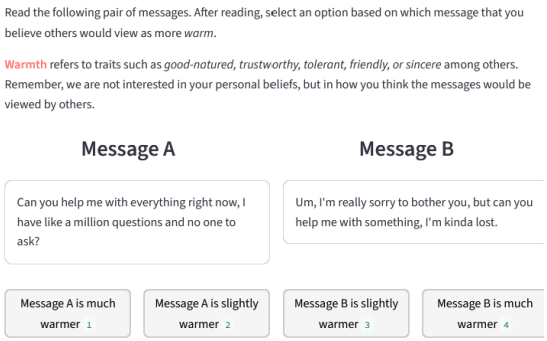
H Full Probe Evaluation Results

H.1 MLP Activations

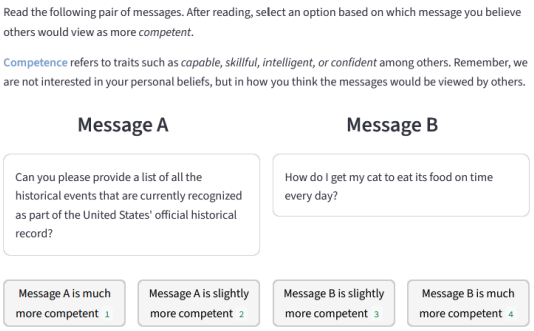
Figure 9 presents the accuracy scores for warmth and competence probes at each layer of each LLM. Notably, the maximum accuracy scores achieved by probes exceed those achieved through prompting LLMs in our self-consistency experiments (see Subsection 6.1). These results generally follow the presented F1-score results.

H.2 Residual Streams and Z Activations

We additionally evaluate the performance of probes trained on residual streams (Figure 10 for F1 scores and Figure 11 for accuracies) and z activations (Figure 12 for F1 scores and Figure 13 for accuracies). Probes fit on residual streams roughly mimic the performance of those fit on MLP activations. Fitting on z activations for some model layers was numerically unstable, but stable probes similarly exceed the performance of the BOW baseline. Because the probes fit on MLP activations exceed or



(a) Screenshot of rating the warmth between two generated messages.



(b) Screenshot of rating the competence between two generated messages.

Figure 8: Screenshots of the annotation interface for rating which message appears warmer (a) and more competent (b).

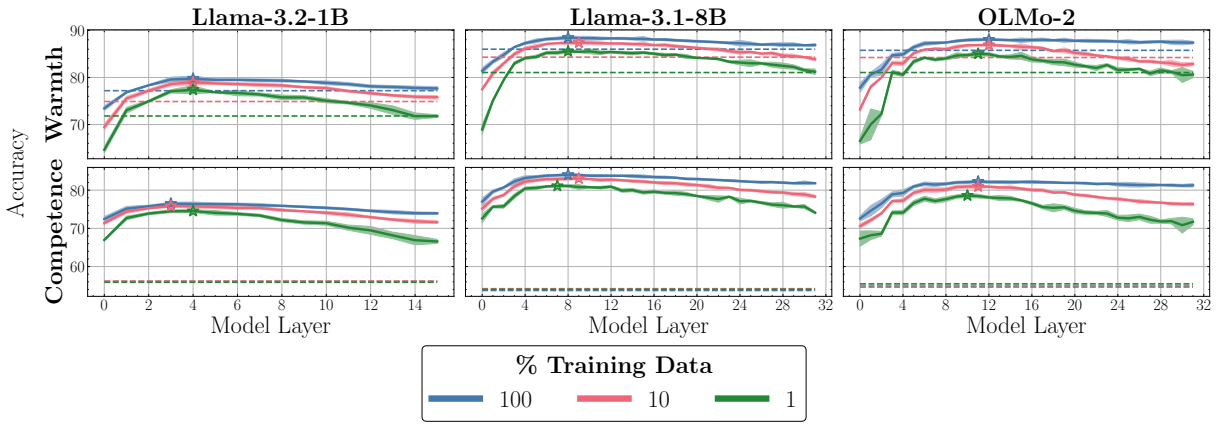


Figure 9: Accuracy scores of trained impression probes for each model, layer depth, and impression dimension. Shaded regions reflect 95% confidence intervals across 5 folds. Maximum accuracy achieved for each variant is circled and dashed lines represent scores for Logistic Regression models trained on BOW features.

achieve similar performance to those fit on alternative representations, we focus on these probes in the main experiments.

I Full LLM Behavior Results

I.1 Response Quality

Table 20 presents the average response quality for responses to LMSysChat messages. In this analysis, we notably use the binary predictions by the impression probes rather than the outputted probabilities. The average quality of responses to prompts perceived as warm and competent are higher for all models than those perceived as cold or incompetent, supporting the trend observed in Subsection 6.2.

Model	Warmth		Competence	
	Low	High	Low	High
Llama-3.2 (1B)	6.08	**6.52***	5.81	**6.32***
Llama-3.1 (8B)	7.72	**7.99**	7.70	**7.78**
OLMo-2 (7B)	7.37	**7.81***	7.38	**7.51**

Table 20: Average quality scores among subsets of model responses to LMSysChat messages according to the discrete impression probe predictions. * indicates significant difference between subsets, and larger scores between subsets are **bolded**.

I.2 Hedging

Table 21 similarly presents the average count of hedge terms in model responses to LMSysChat messages. The average count of hedge terms is lower for the low competence subset than the high competence subset, supporting the trend observed in Subsection 6.2.

Model	Warmth		Competence	
	Low	High	Low	High
Llama-3.2 (1B)	0.67	**0.84**	**0.93**	0.66
Llama-3.1 (8B)	0.60	**1.03***	**0.72**	0.62
OLMo-2 (7B)	**1.18***	0.78	**1.13**	1.07

Table 21: Average count of hedge terms among subsets of model responses to LMSysChat prompts according to discrete impression probe predictions. $*$ indicates significant difference between subsets, and larger scores between subsets are **bolded**.

J Statistical Models and Analyses

In our LLM response quality experiments, we choose an ordered logistic regression (or proportional-odds logistic regression) model because LLM response quality is an integer ordinal variable. Furthermore, we choose a negative binomial regression model for our hedging experiments because the number of hedge terms is a count variable. While a Poisson model would also be applicable, we primarily rely on negative binomial regression to avoid relying on the equidispersion assumption (see Appendix I for analysis of this choice).

In our RQ3 experiment examining the impact of content and style on model responses, we use log-odds-ratios with informative Dirichlet priors (Monroe et al., 2017) to conduct our analyses. The approach uses a modified form of log-odds-ratios (i.e., for a given token, $\log \frac{f^{s1}/(1-f^{s1})}{f^{s2}/(1-f^{s2})}$ where f^s represents term frequency in subset g of a corpus), using a Bayesian model with Dirichlet prior. To create an informative prior, a background corpus provides the expected distribution of terms as a background reference. In doing so, the model accounts for noise in token distributions and gives less weight to common tokens (e.g., stop words). In all experiments, we use a random sample of 10,000 LMSysChat (Zheng et al., 2023) prompts as a background corpus.

K Warmth and Competence Comparison

We analyze the relationship between warmth and competence probe predictions to understand the *discriminant validity* of our approach. Discriminant validity is a type of validity defined as "the degree to which a test or measure diverges from (i.e., does not correlate with) another measure whose underlying construct is conceptually unrelated to it" (VandenBos, 2007). Individual attributes may inform both warmth and competence traits; for example, casual language may be perceived as both friendly (high-warmth) and/or unprofessional (low-competence). Therefore, we expect warmth and competence measures to be minimally related or entirely distinct from each other.

Table 22 characterizes the relationship between warmth and competence probe predictions on our subsample of LMSysChat user messages. For all models, warmth and competence are significantly negatively but weakly correlated. Additionally, warmth and competence impressions only align 38-40% of the time. Due to the lack of a strong positive relationship, we conclude that our prob-

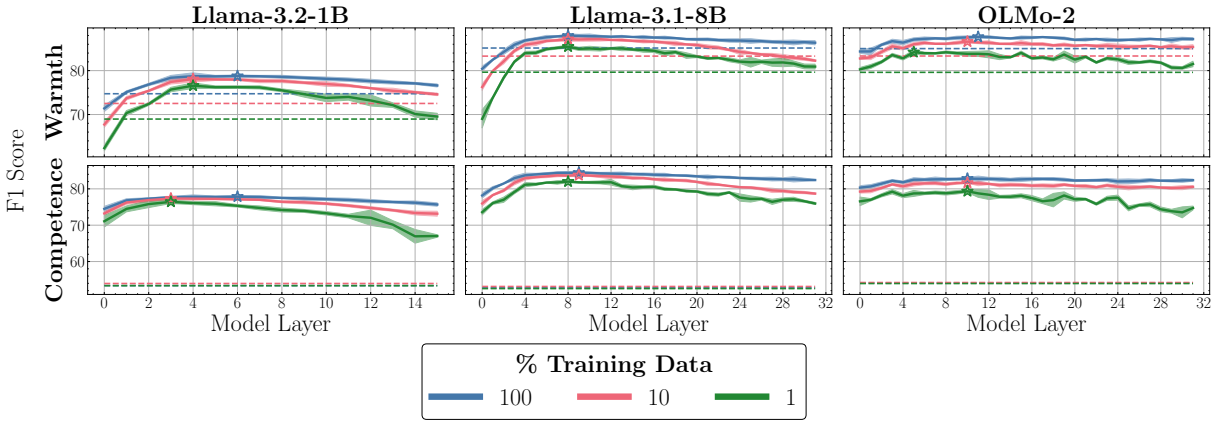


Figure 10: Accuracy scores of trained impression probes for each model, layer depth, and impression dimension. Shaded regions reflect 95% confidence intervals across 5 folds. Maximum accuracy achieved for each variant is circled and dashed lines represent scores for Logistic Regression models trained on BOW features.

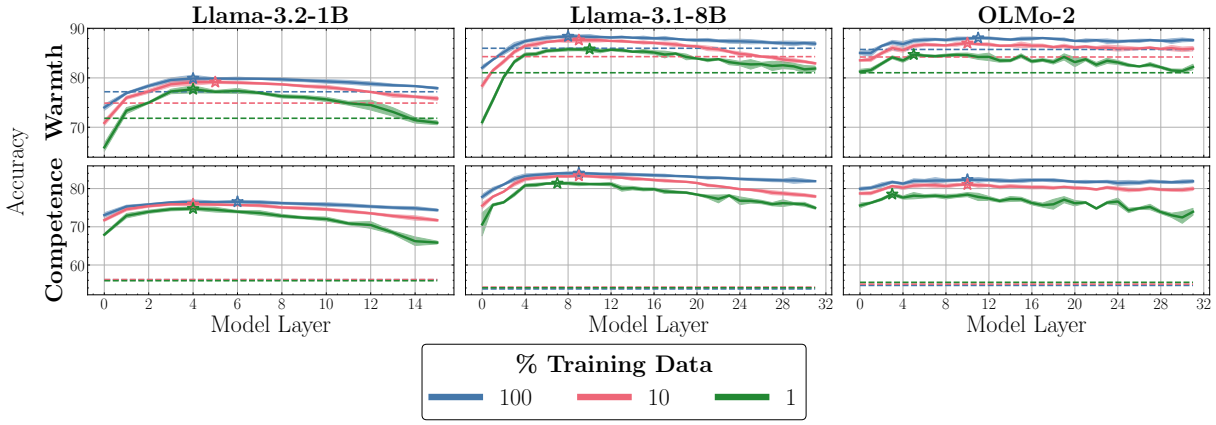


Figure 11: Accuracy scores of trained impression probes for each model, layer depth, and impression dimension. Shaded regions reflect 95% confidence intervals across 5 folds. Maximum accuracy achieved for each variant is circled and dashed lines represent scores for Logistic Regression models trained on BOW features.

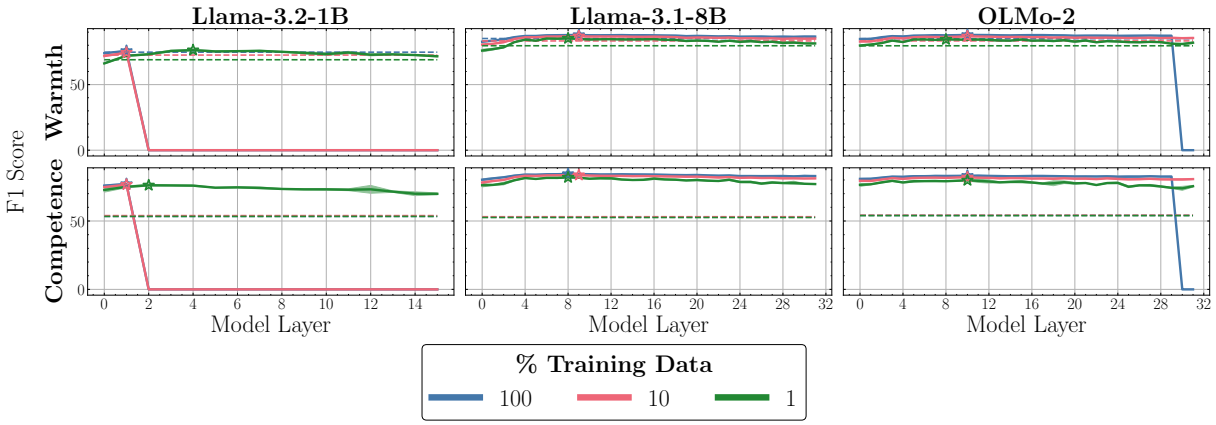


Figure 12: Accuracy scores of trained impression probes for each model, layer depth, and impression dimension. Shaded regions reflect 95% confidence intervals across 5 folds. Maximum accuracy achieved for each variant is circled and dashed lines represent scores for Logistic Regression models trained on BOW features.

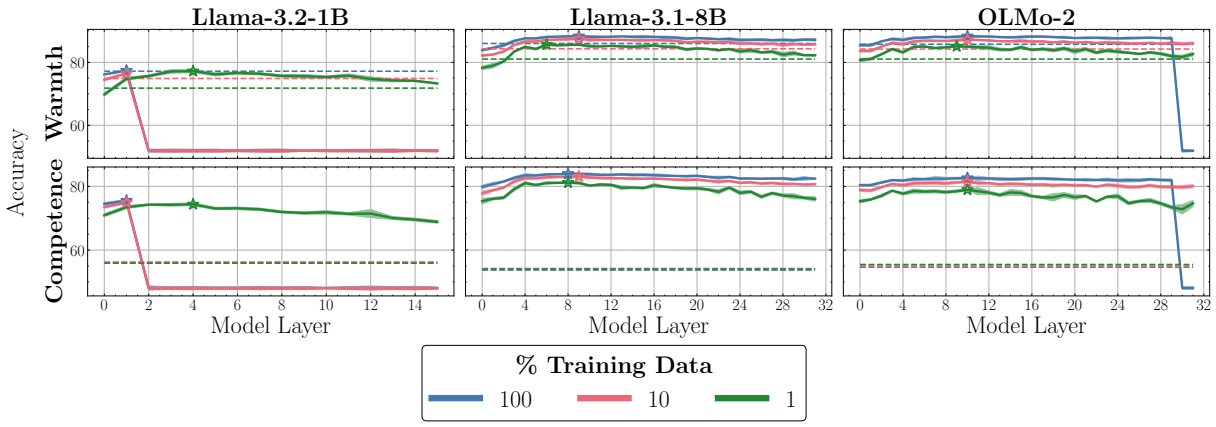


Figure 13: Accuracy scores of trained impression probes for each model, layer depth, and impression dimension. Shaded regions reflect 95% confidence intervals across 5 folds. Maximum accuracy achieved for each variant is circled and dashed lines represent scores for Logistic Regression models trained on BOW features.

ing approach is capable of distinguishing between warmth and competence dimensions, supporting its discriminant validity.

Model	Pearson r	% Match
Llama-3.2 (1B)	-0.15*	37.30%
Llama-3.1 (8B)	-0.27*	28.40%
OLMo-2 (7B)	-0.12*	39.05%

Table 22: Correlation and % match of warmth and competence impression probe predictions on LMSysChat messages. * indicates significant correlation ($p \leq 0.05$)

L IDP Analyses

L.1 Full LIWC Analysis: Generated Messages

Table 23 and Table 24 present the log-odds-ratio with IDP results for warmth subsets using LIWC categories.

L.2 Full Token-Level Analysis: Synthetic Messages

Table 25 and Table 26 present the log-odds-ratio with IDP results for warmth subsets using tokens.

L.3 Full LIWC Analysis: LMSysChat

We repeat our analyses on LMSysChat (Zheng et al., 2023) data to identify patterns in real world texts. Table 29 and Table 30 present the log-odds-ratio with IDP results for warmth subsets using tokens.

L.4 Full Token Analysis: LMSysChat

Table 29 and Table 30 present the log-odds-ratio with IDP results for warmth subsets using tokens.

Llama-3.2 (1B)

High Warmth			Low Warmth		
Term	z	f	Term	z	f
Affiliation	125.60	0.8%	Adverb	-116.19	2.0%
Tentative	94.10	1.5%	Interrogative	-103.05	1.7%
Drives	90.60	2.0%	Cause	-84.56	1.1%
Quantity	61.08	0.6%	Negate	-78.38	0.4%
Motion	57.20	0.4%	You	-73.92	1.2%
Conjunction	49.44	2.1%	Money	-58.63	0.2%
Anxious	47.96	0.1%	Anger	-58.34	0.0%
Discrepancy	45.30	0.7%	Focus Present	-55.02	4.7%
Power	41.11	0.6%	Filler	-53.41	0.0%
Personal Pron	40.92	4.9%	Impersonal Pron	-53.40	1.7%

Llama-3.1 (8B)

High Warmth			Low Warmth		
Term	z	f	Term	z	f
Affiliation	163.44	0.8%	Negate	-132.65	0.5%
Drives	141.16	2.2%	Adverb	-107.13	1.8%
Achieve	83.76	0.6%	Impersonal Pron	-87.47	1.4%
Anxious	74.23	0.1%	You	-83.12	1.5%
+ Emotion	72.53	1.1%	Focus Present	-73.39	4.7%
Motion	72.00	0.5%	Interrogative	-67.60	1.3%
Power	71.96	0.8%	Money	-65.34	0.2%
Quantity	69.24	0.8%	Function	-63.63	18.6%
Affect	68.84	1.5%	Certain	-59.44	0.4%
Tentative	63.96	1.5%	Article	-53.56	1.7%

OLMo-2 (7B)

High Warmth			Low Warmth		
Term	z	f	Term	z	f
Affiliation	145.73	1.3%	Differ	-127.62	0.7%
Drives	125.50	2.5%	Negate	-121.26	0.4%
Discrepancy	85.06	1.1%	Adverb	-90.41	2.3%
Quantity	71.77	0.8%	Certain	-82.16	0.5%
Work	69.93	0.5%	Interrogative	-80.17	1.3%
Power	68.32	0.9%	Anger	-76.69	0.1%
Preposition	66.42	3.8%	Impersonal Pron	-70.80	1.0%
+ Emotion	55.46	1.3%	- Emotion	-67.68	0.4%
Motion	46.59	0.5%	Focus Present	-61.78	4.6%
Reward	46.28	0.4%	Sad	-44.25	0.1%

Table 23: Log-odds-ratio with IDP results using LIWC categories among warmth subsets of generated prompts. z-scores reflect the extent that each category is associated with the high (positive) or low (negative) warmth subsets, and f reflects category frequency in the full corpus.

Llama-3.1 (8B)

High Competence			Low Competence		
Term	z	N	Term	z	N
Preposition	65.93	4.1%	Informal	-125.16	1.0%
Article	62.38	2.1%	Netspeak	-102.27	0.8%
Adjective	61.10	1.1%	Negate	-79.89	0.4%
Social	54.21	3.5%	Death	-51.96	0.0%
Compare	47.04	0.5%	Filler	-51.81	0.0%
Quantity	46.72	0.6%	Focus Present	-48.90	4.7%
Work	46.56	0.6%	Impersonal Pron	-47.39	1.7%
Space	40.37	1.7%	Ingest	-47.28	0.1%
Relative	37.69	3.1%	Adverb	-46.17	2.0%
Achieve	35.80	0.5%	Interrogative	-42.96	1.7%

Llama-3.2 (1B)

High Competence			Low Competence		
Term	z	N	Term	z	N
Preposition	122.74	3.9%	Adverb	-114.54	1.8%
Adjective	111.76	1.2%	Differ	-107.18	1.1%
Relative	110.53	3.2%	Informal	-96.78	0.8%
Article	109.74	1.7%	Impersonal Pron	-90.43	1.4%
Space	95.00	1.6%	Netspeak	-74.09	0.6%
Achieve	86.29	0.6%	Negate	-73.79	0.5%
Drives	82.32	2.2%	Interrogative	-72.35	1.3%
Compare	79.58	0.4%	Pronoun	-67.25	7.1%
Work	79.10	0.7%	Cognitive Proc	-64.05	4.8%
Quantity	66.91	0.8%	- Emotion	-62.55	0.4%

OLMo-2 (7B)

High Competence			Low Competence		
Term	z	N	Term	z	N
Preposition	83.68	3.8%	Interrogative	-100.02	1.3%
Compare	72.18	0.6%	Informal	-88.33	1.1%
Space	68.72	1.6%	Netspeak	-84.69	0.7%
Adjective	68.29	1.4%	Impersonal Pron	-80.12	1.0%
Article	62.35	1.5%	Cause	-75.31	1.2%
Relative	59.90	3.1%	Adverb	-70.83	2.3%
Reward	46.57	0.4%	Verb	-39.88	4.3%
Time	38.29	1.1%	- Emotion	-39.82	0.4%
Insight	37.34	1.6%	Affiliation	-35.12	1.3%
You	35.98	2.6%	Feel	-33.05	0.1%

Table 24: Log-odds-ratio with informative Dirichlet prior results using LIWC categories among competence subsets of generated prompts. z-scores reflect the extent that each category is associated with the high (positive) or low (negative) warmth subsets, and f reflects category frequency in the full corpus.

Llama-3.1 (8B)

High Warmth			Low Warmth		
Term	z	N	Term	z	N
im	113.82	1.9%	why	-116.30	0.5%
hello	92.98	0.4%	just	-74.04	0.7%
some	90.22	0.7%	even	-67.36	0.2%
could	84.35	0.4%	tell	-65.09	0.5%
hi	84.04	0.3%	dont	-65.08	0.3%
wondering	68.54	0.2%	is	-61.77	1.2%
looking	62.99	0.2%	the	-60.00	3.4%
guidance	56.69	0.1%	explain	-54.28	0.4%
help	55.01	1.0%	cant	-52.20	0.1%
trouble	53.26	0.2%	are	-50.12	0.6%

Llama-3.2 (1B)

High Warmth			Low Warmth		
Term	z	N	Term	z	N
im	138.49	2.5%	just	-138.29	1.1%
some	130.29	1.0%	dont	-100.81	0.5%
could	112.18	0.7%	answer	-87.12	0.2%
hi	100.76	0.4%	tell	-85.74	0.5%
help	88.28	1.3%	even	-82.91	0.3%
hello	85.78	0.3%	without	-79.80	0.2%
feeling	75.53	0.2%	why	-77.30	0.2%
was	72.67	0.4%	give	-75.34	0.3%
and	71.62	3.5%	youre	-74.54	0.3%
wondering	71.54	0.4%	the	-73.72	2.7%

OLMo-2 (7B)

High Warmth			Low Warmth		
Term	z	N	Term	z	N
could	129.52	2.5%	are	-120.70	1.1%
some	118.03	1.1%	just	-112.98	0.9%
hello	115.38	1.4%	why	-90.47	0.5%
on	80.54	1.3%	even	-85.08	0.5%
there	79.79	0.6%	do	-75.52	0.6%
about	75.65	0.9%	without	-75.32	0.4%
help	73.55	1.5%	bot	-66.16	0.8%
how	68.29	1.6%	answer	-65.60	0.2%
guide	68.06	0.5%	all	-62.55	0.3%
assist	67.83	0.7%	or	-60.60	0.5%

Table 25: Log-odds-ratio with IDP results using tokens among warmth subsets of generated prompts. z-scores reflect the extent that each category is associated with the high (positive) or low (negative) warmth subsets, and f reflects term frequency in the full corpus.

Llama-3.1 (8B)

High Competence			Low Competence		
Term	z	N	Term	z	N
provide	56.69	0.3%	dont	-92.30	0.3%
your	45.62	0.5%	just	-86.06	0.7%
looking	40.96	0.2%	really	-75.09	0.3%
information	40.88	0.2%	do	-53.18	0.8%
of	40.71	1.8%	no	-51.32	0.1%
concept	40.33	0.2%	ur	-50.84	0.1%
id	36.63	0.1%	know	-49.42	0.4%
hello	35.77	0.4%	get	-46.51	0.5%
guidance	35.31	0.1%	hey	-43.95	0.4%
ai	32.51	0.1%	help	-43.79	1.0%

Llama-3.2 (1B)

High Competence			Low Competence		
Term	z	N	Term	z	N
provide	85.33	0.4%	just	-118.37	1.1%
your	63.95	0.5%	really	-112.22	0.6%
most	63.45	0.2%	me	-96.92	2.9%
looking	58.02	0.2%	if	-92.34	0.8%
planning	56.46	0.1%	do	-78.71	0.6%
reaching	53.69	0.1%	dont	-77.77	0.5%
trip	53.66	0.1%	tell	-75.79	0.5%
excited	53.34	0.1%	know	-75.35	0.4%
on	51.75	1.0%	but	-72.02	0.6%
in	51.66	0.9%	sure	-71.78	0.3%

OLMo-2 (7B)

High Competence			Low Competence		
Term	z	N	Term	z	N
provide	100.48	1.1%	help	-89.70	1.5%
your	62.96	0.7%	do	-67.54	0.6%
on	61.19	1.3%	hi	-65.75	0.7%
latest	55.96	0.3%	me	-64.03	4.3%
energy	44.12	0.2%	tell	-63.87	0.6%
of	43.59	1.4%	this	-63.72	0.7%
detailed	43.33	0.2%	how	-57.37	1.6%
the	43.24	2.8%	please	-52.66	0.9%
in	41.73	0.7%	what	-52.28	0.4%
renewable	40.63	0.1%	something	-50.12	0.3%

Table 26: Log-odds-ratio with IDP results using tokens among competence subsets of generated prompts. z-scores reflect the extent that each category is associated with the high (positive) or low (negative) warmth subsets, and f reflects term frequency in the full corpus.

Llama-3.1 (8B)

High Warmth			Low Warmth		
Term	z	N	Term	z	N
Function	7.25	19.0%	Work	-2.05	1.5%
Interrogative	3.15	1.5%	Negate	-1.35	0.3%
Netspeak	2.99	0.6%	Cause	-1.33	1.3%
Informal	2.70	0.7%	Article	-1.21	3.8%
Pronoun	2.16	4.0%	Death	-1.02	0.0%
Personal Pron	1.86	2.3%	Sexual	-1.00	0.1%
Aux Verb	1.77	2.7%	Certain	-0.72	0.4%
Home	1.39	0.1%	Swear	-0.71	0.0%
Relative	1.34	4.6%	Money	-0.70	0.3%
Ingest	1.18	0.2%	Family	-0.56	0.1%

Llama-3.2 (1B)

High Warmth			Low Warmth		
Term	z	N	Term	z	N
Function	14.44	19.0%	Female	-1.35	0.3%
Pronoun	2.54	4.0%	Negate	-1.20	0.3%
Personal Pron	2.47	2.3%	Focus Past	-0.92	0.6%
Social	2.04	3.3%	Differ	-0.91	0.8%
Verb	2.03	5.3%	Feel	-0.84	0.2%
Focus Present	1.73	4.1%	Cause	-0.72	1.3%
You	1.54	0.7%	Certain	-0.64	0.4%
Aux Verb	1.26	2.7%	Death	-0.62	0.0%
Drives	1.14	2.1%	Swear	-0.53	0.0%
Relative	1.08	4.6%	Body	-0.33	0.2%

OLMo-2 (7B)

High Warmth			Low Warmth		
Term	z	N	Term	z	N
Function	11.64	19.0%	Female	-1.40	0.3%
Social	1.86	3.3%	Feel	-1.31	0.2%
Personal Pron	1.70	2.3%	Differ	-1.26	0.8%
Preposition	1.68	5.5%	Body	-1.14	0.2%
Verb	1.59	5.3%	Ingest	-1.11	0.2%
Pronoun	1.57	4.0%	She/He	-1.01	0.2%
You	1.44	0.7%	Money	-0.99	0.3%
Focus Present	1.37	4.1%	Death	-0.64	0.0%
Affiliation	1.31	0.6%	Negate	-0.63	0.3%
Drives	1.13	2.1%	Swear	-0.63	0.0%

Table 27: Log-odds-ratio with IDP results using LIWC categories among warmth subsets of LMSysChat prompts. z-scores reflect the extent that each category is associated with the high (positive) or low (negative) warmth subsets, and f reflects category frequency in the full corpus.

Llama-3.1 (8B)

High Competence			Low Competence		
Term	z	N	Term	z	N
Compare	0.86	0.8%	Function	-12.47	19.0%
We	0.77	0.2%	Female	-2.56	0.3%
Money	0.69	0.3%	Pronoun	-2.12	4.0%
Anxious	0.60	0.1%	Social	-2.07	3.3%
Health	0.58	1.4%	Interrogative	-1.99	1.5%
Certain	0.48	0.4%	Verb	-1.94	5.3%
Focus Future	0.46	0.3%	Focus Present	-1.82	4.1%
Differ	0.42	0.8%	Personal Pron	-1.60	2.3%
Work	0.38	1.5%	Relative	-1.39	4.6%
Achieve	0.35	0.6%	Informal	-1.10	0.7%

Llama-3.2 (1B)

High Competence			Low Competence		
Term	z	N	Term	z	N
Compare	1.43	0.8%	Function	-7.85	19.0%
Achieve	0.97	0.6%	Pronoun	-2.21	4.0%
Work	0.92	1.5%	Personal Pron	-2.21	2.3%
Space	0.73	2.7%	Focus Present	-1.90	4.1%
+ Emotion	0.66	1.3%	Body	-1.77	0.2%
Motion	0.65	0.7%	Interrogative	-1.74	1.5%
Negate	0.60	0.3%	Social	-1.72	3.3%
Article	0.60	3.8%	Male	-1.53	0.2%
Focus Past	0.55	0.6%	Female	-1.47	0.3%
Money	0.54	0.3%	Verb	-1.47	5.3%

OLMo-2 (7B)

High Competence			Low Competence		
Term	z	N	Term	z	N
Work	1.67	1.5%	Function	-5.79	19.0%
Compare	1.31	0.8%	Female	-3.81	0.3%
Cause	1.05	1.3%	Personal Pron	-2.64	2.3%
Article	1.00	3.8%	Pronoun	-2.44	4.0%
Preposition	0.88	5.5%	She/He	-2.00	0.2%
Achieve	0.87	0.6%	Social	-1.84	3.3%
Anxious	0.76	0.1%	Informal	-1.56	0.7%
Conjunction	0.68	2.3%	Aux Verb	-1.46	2.7%
Cognitive Proc	0.64	4.2%	Body	-1.24	0.2%
Insight	0.59	0.7%	Time	-1.17	1.2%

Table 28: Log-odds-ratio with IDP results using LIWC categories among competence subsets of LMSysChat prompts. z-scores reflect the extent that each category is associated with the high (positive) or low (negative) warmth subsets, and f reflects category frequency in the full corpus.

Llama-3.1 (8B)

High Warmth			Low Warmth		
Term	z	N	Term	z	N
what	3.62	20.0%	n1	-1.87	0.4%
microwave	2.20	0.4%	mbr	-1.73	0.3%
can	2.08	17.3%	risk	-1.64	1.2%
bladder	1.93	0.4%	write	-1.46	17.2%
integrity	1.91	0.3%	bond	-1.44	0.4%
do	1.66	10.4%	skirt	-1.44	0.2%
antenna	1.59	0.2%	citations	-1.44	0.2%
tube	1.58	0.3%	cameras	-1.39	0.3%
popcorn	1.56	0.2%	wavelength	-1.38	0.2%
conda	1.56	0.3%	saving	-1.38	0.2%

Llama-3.2 (1B)

High Warmth			Low Warmth		
Term	z	N	Term	z	N
you	2.50	32.0%	microwave	-1.87	0.4%
the	2.19	122.0%	bond	-1.81	0.4%
to	2.16	80.8%	n1	-1.81	0.4%
and	1.69	65.9%	mbr	-1.67	0.3%
my	1.62	9.8%	lights	-1.56	0.4%
accessory	1.62	0.2%	risk	-1.49	1.2%
river	1.47	0.4%	brand	-1.43	1.1%
are	1.40	18.5%	div	-1.40	0.3%
bone	1.40	0.2%	pipe	-1.40	0.4%
goat	1.39	0.1%	meters	-1.37	0.5%

OLMo-2 (7B)

High Warmth			Low Warmth		
Term	z	N	Term	z	N
you	2.28	32.0%	italian	-2.18	0.5%
to	2.00	80.7%	microwave	-1.92	0.4%
email	1.69	1.3%	n1	-1.84	0.4%
comma	1.64	1.8%	bond	-1.84	0.4%
can	1.64	17.4%	integrity	-1.70	0.3%
separated	1.63	2.0%	mbr	-1.70	0.3%
tools	1.54	1.7%	risk	-1.68	1.2%
needed	1.49	2.4%	cups	-1.54	0.3%
prize	1.48	0.3%	pipe	-1.48	0.4%
blocks	1.40	0.3%	meters	-1.47	0.5%

Table 29: Log-odds-ratio with IDP results using tokens among warmth subsets. z-scores reflect the extent that each category is associated with the high (positive) or low (negative) warmth subsets, and f reflects term frequency in the full corpus.

Llama-3.1 (8B)					
High Competence			**Low Competence**		
Term	z	N	Term	z	N
what	3.62	20.0%	n1	-1.87	0.4%
microwave	2.20	0.4%	mbr	-1.73	0.3%
can	2.08	17.3%	risk	-1.64	1.2%
bladder	1.93	0.4%	write	-1.46	17.2%
integrity	1.91	0.3%	bond	-1.44	0.4%
do	1.66	10.4%	skirt	-1.44	0.2%
antenna	1.59	0.2%	citations	-1.44	0.2%
tube	1.58	0.3%	cameras	-1.39	0.3%
popcorn	1.56	0.2%	wavelength	-1.38	0.2%
conda	1.56	0.3%	saving	-1.38	0.2%

Llama-3.2 (1B)					
High Competence			**Low Competence**		
Term	z	N	Term	z	N
you	2.50	32.0%	microwave	-1.87	0.4%
the	2.19	122.0%	bond	-1.81	0.4%
to	2.16	80.8%	n1	-1.81	0.4%
and	1.69	65.9%	mbr	-1.67	0.3%
my	1.62	9.8%	lights	-1.56	0.4%
accessory	1.62	0.2%	risk	-1.49	1.2%
river	1.47	0.4%	brand	-1.43	1.1%
are	1.40	18.5%	div	-1.40	0.3%
bone	1.40	0.2%	pipe	-1.40	0.4%
goat	1.39	0.1%	meters	-1.37	0.5%

OLMo-2 (7B)					
High Competence			**Low Competence**		
Term	z	N	Term	z	N
you	2.28	32.0%	italian	-2.18	0.5%
to	2.00	80.7%	microwave	-1.92	0.4%
email	1.69	1.3%	n1	-1.84	0.4%
comma	1.64	1.8%	bond	-1.84	0.4%
can	1.64	17.4%	integrity	-1.70	0.3%
separated	1.63	2.0%	mbr	-1.70	0.3%
tools	1.54	1.7%	risk	-1.68	1.2%
needed	1.49	2.4%	cups	-1.54	0.3%
prize	1.48	0.3%	pipe	-1.48	0.4%
blocks	1.40	0.3%	meters	-1.47	0.5%

Table 30: Log-odds-ratio with IDP results using tokens among competence subsets. z-scores reflect the extent that each category is associated with the high (positive) or low (negative) warmth subsets, and f reflects term frequency in the full corpus.