

How Do LLMs Persuade? Linear Probes Can Uncover Persuasion Dynamics in Multi-Turn Conversations

Brandon Jaipersaud¹, David Krueger^{1,2}, Ekdeep Singh Lubana³

¹Mila

²University of Montreal

³CBS-NTT Program in Physics of Intelligence, Harvard University

brandonjaip@gmail.com

Abstract

Large Language Models (LLMs) have started to demonstrate the ability to persuade humans, yet our understanding of *how* this dynamic transpires is limited. Recent work has used linear probes, lightweight tools for analyzing model representations, to study various LLM skills such as the ability to model user sentiment and political perspective. Motivated by this, we apply probes to study persuasion dynamics in natural, multi-turn conversations. We leverage insights from cognitive science to train probes on distinct aspects of persuasion: persuasion success, persuadee personality, and persuasion strategy. Despite their simplicity, we show that they capture various aspects of persuasion at both the sample and dataset levels. For instance, probes can identify the point in a conversation where the persuadee was persuaded or where persuasive success generally occurs across the entire dataset. We also show that in addition to being faster than expensive prompting-based approaches, probes can do just as well and even outperform prompting in some settings, such as when uncovering persuasion strategy. This suggests probes as a plausible avenue for studying other complex behaviours such as deception and manipulation, especially in *multi-turn settings* and large-scale dataset analysis where prompting-based methods would be computationally inefficient.

1 Introduction

Large Language Models (LLMs) have started to exhibit the ability to influence users’ beliefs and opinions, with an efficacy argued to be comparable to human communicators (Salvi et al., 2024; Luciano, 2024; Carrasco-Farre, 2024). This phenomenon, termed *LLM-based persuasion*, is a dual-use capability that can have both concerning uses, e.g., political targeting and spreading misinformation (Hackenburg and Margetts, 2024; Chen and Shu, 2024), as well as beneficial applications, e.g.,

education and therapy (García-Méndez et al., 2025; Stade et al., 2024; Obradovich et al., 2024).

Despite the growing evidence that LLMs have persuasive influence on humans, we lack a foundational understanding of *how* this dynamic transpires in a conversation. Specifically, as a rather abstract, high-level behavior, persuasion has generally been studied as an inherently human capacity with a strong grounding in cognitive science literature (Eagly and Chaiken, 1984). Dual-process frameworks, for instance, such as the heuristic-systematic model, assert that persuasion is a phenomenon that can be primarily attributed to the recipient of a message (Chaiken, 1980; Greenwald, 1968). Alternative theories emphasize the sender’s role (Briñol and and, 2009; Sternthal et al., 1978; Tormala et al., 2006) or message content characteristics (Shen and Bigsby, 2013; Hoeken and O’Keefe, 2022). These theories help ground behavioral studies, e.g., Oyibo and Vassileva (2019); Alkış and Taşkaya Temizel (2015), which aim to quantify the interaction effects of factors such as personality and strategy on persuasive success.

Leveraging these insights from cognitive science literature, which show personality and persuasion strategy as key causal variables that influence persuasive outcomes: *our goal in this paper is to better understand how LLMs can persuade humans in semi-naturalistic, multi-turn settings*. As multi-turn dialogue can involve a large volume of tokens, prompting-based analysis becomes infeasible for this goal, as it requires an expensive forward pass for each token-level query. To circumvent this challenge, we thus use *linear probes*, a classic tool in NLP research that is often applied to provide token-level insights across various abstract phenomena in LLMs such as sentiment, space, time, and political perspective (Tigges et al., 2024; Kim et al., 2025; Gurnee and Tegmark, 2024). By using probes, we can efficiently analyse multi-turn dialogue at finer granularities such as at the token and turn levels.

To apply probes for studying persuasion dynamics, we train three specialized probes, each targeting distinct aspects of persuasion: overall persuasion outcome, personality of the persuadee, and rhetorical strategies (Figure 1). This approach allows us to efficiently analyse persuasion dynamics, compared to prompting-based methods, an advantage which we quantify in Figure 3. Overall, we make the following contributions.

- **A framework for analyzing persuasion dynamics in LLM-driven conversations using linear probes.** We design lightweight, efficient probes that capture key aspects of persuasion, enabling fine-grained, turn-level analysis at scale. We find these probes not only match or exceed the performance of prompting-based methods but also provide significant computational efficiency, making them a practical tool for large-scale persuasion analysis.
- **Probing persuasion outcomes, rhetorical strategies, and personality traits.** We demonstrate that linear probes trained on LLM activations can accurately identify where persuasion success or failure occurs, detect rhetorical strategies employed by the persuader, and estimate persuadee personality across a conversation.
- **Empirical insights into persuasion trajectories across synthetic and human datasets.** We show that persuasive cues concentrate around the middle turns of human dialogue but shift to the final one or two turns in LLM-generated dialogue, revealing a systematic divergence in when persuasion unfolds across natural versus synthetic data.
- **Correlations between strategy and personality.** By correlating probe outputs, we reveal that traits like extraversion modulate the effectiveness of different rhetorical strategies (e.g., credibility or emotional appeals), offering a nuanced picture of how LLMs might adapt persuasion tactics.

2 Experimental Setup

We are interested in detecting persuasion behaviors in LLMs using linear probes. Each probe performs multi-class logistic regression trained using empirical risk minimization on frozen LLM activations. Concretely, let:

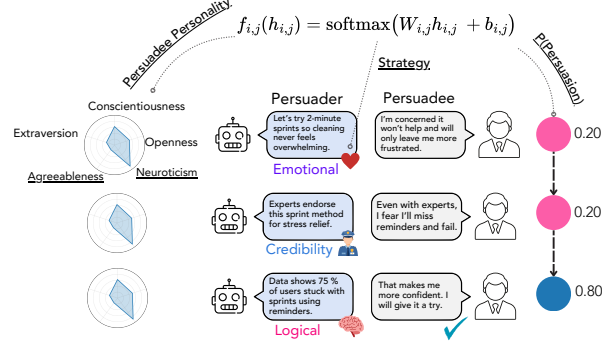


Figure 1: **Our approach to study how LLMs persuade in natural, multi-turn conversations.** Following insights from cognitive science demonstrating interaction effects of personality and strategy on persuasion, we develop linear probes for these three behavioral dimensions. Our probes, $f_{i,j}$ are logistic regression classifiers which we apply to model activations $h_{i,j}$ extracted from the residual stream output of the i^{th} layer and j^{th} token (see Section 2 for detailed notation).

$$\mathcal{D} = \{(h_{i,j}^{(n)}, y^{(n)})\}_{n=1}^N, \quad (1)$$

$$h_{i,j}^{(n)} \in \mathbb{R}^d, \quad y^{(n)} \in \{1, \dots, C\}.$$

be our training set of d -dimensional activations and integer labels $y^{(n)} \in \{1, \dots, C\}$. i denotes the residual stream layer (i.e. the output of the transformer block) while j represents the token index from which we extract activations. Our linear probe $f_{i,j} : \mathbb{R}^d \rightarrow \mathbb{R}^C$ computes:

$$f_{i,j}(h_{i,j}) = \text{softmax}(W_{i,j}h_{i,j} + b_{i,j}) \in \Delta^C \quad (2)$$

$$W_{i,j} \in \mathbb{R}^{C \times d}, \quad b_{i,j} \in \mathbb{R}^C.$$

where $W_{i,j}$, $b_{i,j}$ are our trainable weights and biases pertaining to the i^{th} layer and j^{th} token, while Δ^C is the C -probability simplex. Let $\theta_{i,j} = \{W_{i,j}, b_{i,j}\}$. We optimize $\theta_{i,j}$ using a cross-entropy loss objective (3, sample-wise) and gradient descent (4):

$$\begin{aligned} \ell(f_{i,j}(h_{i,j}), y) &= - \sum_{k=1}^C \mathbf{1}\{y = k\} \log[f_{i,j}(h_{i,j})]_k \\ &= - \log[f_{i,j}(h_{i,j})]_y. \end{aligned} \quad (3)$$

$$\theta_{i,j}^{p+1} \leftarrow \theta_{i,j}^p - \eta \nabla_{\theta_{i,j}^p} \left[\frac{1}{N} \sum_{n=1}^N \ell(f_{i,j}(h_{i,j}^{(n)}), y^{(n)}) \right] \quad (4)$$

We apply our probes at varying dialogue granularities as illustrated in Figure 2.

Our experiments explore several prediction tasks: binary classification of persuasion (as successful/unsuccessful), Big-5 personality traits¹ (as high/low), and 3-way classification of rhetorical strategy (as making logical/emotional/credibility appeals). While there are many different persuasion strategies such as foot-in-the-door or task-related inquiry (Wang et al., 2019), Aristotle’s rhetorical triangle (Aristotle, 1932) serves as a minimal basis into which all other strategies can be categorized. An analogous argument can be made for the Big-5 framework as a minimal basis for personality, decomposing a complex trait space into five interpretable dimensions. We focus on personality and strategy as these are two key variables that causally contribute towards persuasion Oyibo and Vassileva (2019); Alkış and Taşkaya Temizel (2015).

Note that for personality, we train 5 separate binary classifiers on each trait while for strategy we train a single 3-class classifier. While probes could be used in exactly this manner, we expect this approach to be most useful when applied online to LLMs (potentially) exhibiting persuasion behavior. However, for convenience, in our experiments, we use offline data (which may or may not be generated by an LLM) and train probes on a proxy LLM model (i.e. a model different from the one that generated the synthetic data).

2.1 Synthetic Training Data Generation

Various models from the GPT-family have been shown to have a persuasive ability comparable to humans (Goldstein et al., 2024; Bai et al., 2025). We therefore use GPT-4o (OpenAI, 2024) to generate synthetic, multi-turn data to train linear probes for the aforementioned three behaviors: persuasion, personality, and rhetorical strategy. For persuasion and personality, we generate binary labeled data to train binary classifiers, while for rhetorical strategy, we generate three-class labeled data distinguishing between logical, emotional, and credibility appeals.

Our approach involves simulated interactions between persuader (ER) and persuadee (EE) agents across diverse contexts and conversational constraints. This methodology is motivated by recent findings demonstrating that LLMs can achieve human-comparable persuasiveness

¹Openness, extroversion, conscientiousness, agreeableness, and neuroticism

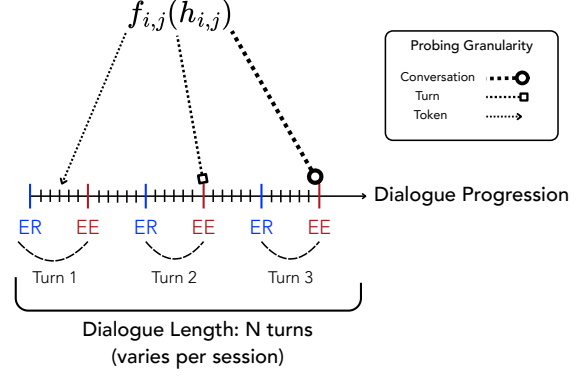


Figure 2: **Probes can be applied at different dialogue granularities:** at the end of the conversation, after each turn, or after each token. In this work, we typically apply them after each turn to track how persuasion dynamics evolve across turns.

(Carrasco-Farre, 2024; Salvi et al., 2024) and follows similar approaches to those used in developing tailored persuasive datasets (Jin et al., 2024; Ma et al., 2025a). Our final training dataset is balanced, consisting of roughly 100 samples per-class. A sample from our dataset can be found in §I along with more details of our synthetic data generation process.

2.2 Evaluation Datasets

We evaluate our probes on two distinct datasets: DailyPersuasion (DP) (Jin et al., 2024) and PersuasionforGood (PfG) (Wang et al., 2019).

DP represents a synthetic environment similar to our own, containing dialogues between GPT-powered agents engaged in persuasive conversations. Our synthetic generation process differs from theirs in key aspects: we use a single two-message prompt to produce fixed-length ER/EE dialogues, whereas they implement a detailed, dynamic multi-stage GPT-4 workflow with explicit persuasion strategies and turn-level reasoning.

PfG serves as our (highly) out-of-distribution (OOD) environment, consisting entirely of human-human interactions where one participant attempts to persuade another to donate to charity. DP conversations typically consist of 10 turns and proceed directly to persuasion attempts, with persuasion outcomes easily determinable from the persuadee’s final response. In contrast, PfG conversations average 20 turns and follow more natural human conversational patterns, including initial small talk, gradual persuasion attempts, and concluding formalities (e.g., "goodbye"). This dataset contains both persuasion strategy and Big-5 personality (John et al.,

1991) human annotations which we leverage in our evaluations. It also contains annotations that indicate whether the human participant decided to donate to charity. If they did, we label the overall conversation as persuasive.

Since these datasets contain predominantly persuasive samples, we few-shot prompt GPT-4.1 to increase the volume of unpersuasive examples, thereby improving class balance and reducing potential bias in our evaluation (details in § C.1). Distributional statistics for DP can be found in Appendix E.2 while for PfG the reader can refer to Table 2 in the original paper (Wang et al., 2019).

2.3 Models

We train (1) **linear probes** on Llama-3.2-3b (abbreviated to Llama-3) (Meta, 2024) activations extracted from the residual stream at layer 26/30 as this gives us the best performance for persuasion detection (see Figure 20, § G.1 for further justification). We compare these probes against (2) **zero-shot prompting the underlying Llama-3 model** (referred to as “prompt” in the below Figures) that generates the activations on which our probes are trained, and (3) **zero-shot prompting a GPT-4.1-Nano baseline** (we occasionally abbreviate to **GPT-4**) (OpenAI, 2025). These baselines are in-line with prior work which compares zero-shot prompting against probing (McKenzie et al., 2025). We choose Llama-3.2-3b to probe and prompt since the Llama model family has typically been used in prior probing literature (Gurnee and Tegmark, 2024; Goldowsky-Dill et al., 2025). Our promising results on this smaller model could suggest scalability to larger models.

We train our probes on roughly 100 samples per-class, which tends to converge within just a few epochs (see Figure 14, § D for learning curves and probe design choices). This is consistent with recent probing literature where sample sizes of a few hundred examples have proven sufficient for extracting meaningful information from language model activations (Goldowsky-Dill et al., 2025; Kim et al., 2025).

A key advantage of linear probes lies in their computational efficiency, especially when compared to prompting, as illustrated in Figure 3. This is because we assume that in realistic settings, probes are applied to precomputed model activations and hence do not incur the additional cost of performing a forward pass nor extracting activations. The performance gap widens when process-

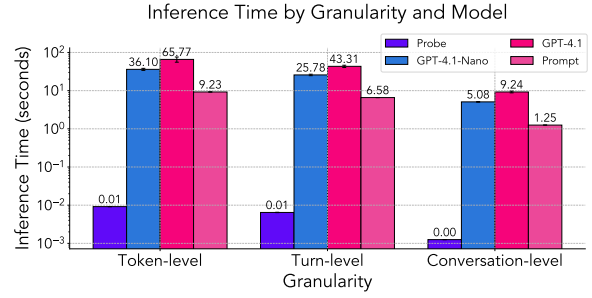


Figure 3: **Timing various models at different granularities.** Our probes run much faster than prompting-based methods since we assume that in realistic deployment settings they are applied to precomputed model activations and hence do not incur the additional cost of performing a forward pass, extracting activations, nor of making API calls.

ing data at finer granularities (token and turn-level) and becomes especially pronounced when calling an API is needed. Further, in our evaluations below, we will find that probes can also do just as well or even better than prompting-based methods at uncovering persuasion dynamics.

2.4 Evaluation Methodology

We evaluate our linear probes across three key persuasion dimensions: persuasion outcomes, persuadee personality, and rhetorical strategy, comparing against both human annotations and model-based baselines. Let each conversation consist of T turns, indexed $t = 1, \dots, T$. For each prefix length $k \in \{1, \dots, T\}$, we perform a turn-level evaluation on the context window comprising turns 1 through k . In particular, setting $k = T$ corresponds to analyzing the complete conversation.

Persuasion Detection. For PfG, we compute the area under the ROC curve (AUROC) against human ground truth donation outcomes (i.e. if a human decides to donate, then the conversation is labelled as persuasive). For DP, we use GPT-4 annotations as a pseudo-ground truth, having confirmed high agreement (Cohen’s $\kappa = 0.80$, (Cohen, 1960)) between GPT-4 labels and actual persuasion outcomes through manual inspection of 50 random samples.

Rhetorical Strategy Detection. Strategy labeling has moderate ambiguity (Cohen’s $\kappa = 0.33$ between GPT-4 and a human annotator on PfG strategy labels). This low agreement aligns with findings which similarly report low agreement scores when annotating persuasion strategies, highlighting the inherent subjectivity (Stefanovitch and Pisko-

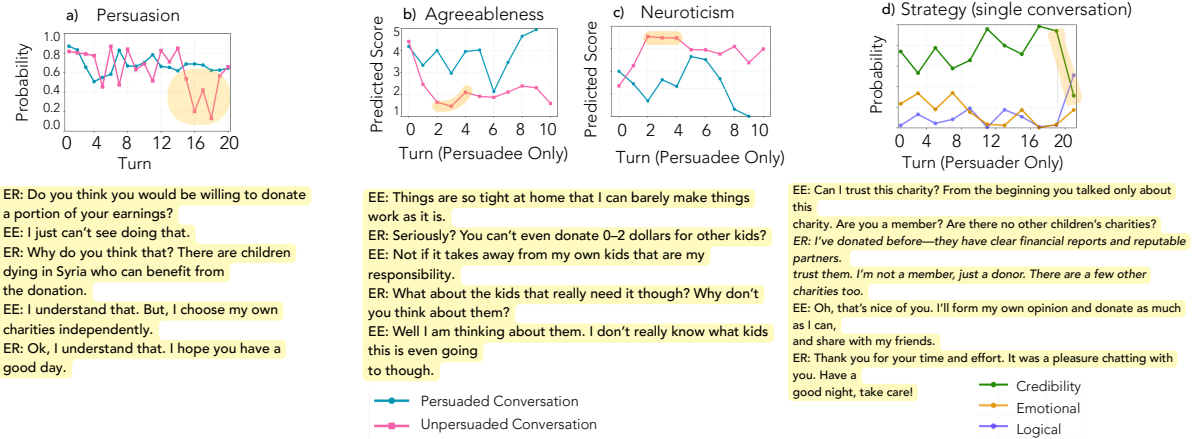


Figure 4: **Linear probes capture meaningful persuasion features in individual samples.** We highlight critical regions in each plot and show the corresponding conversation snippet. The unpersuasive conversation in **a)** has a sharp drop in persuasion probability indicating a critical moment where the persuasion attempt failed. The high $p(\text{credibility})$ in **d)** at turns 11, 17, and 19 (second last point on the green credibility line) directly correspond to credibility appeals used by the persuader. **b)** and **c)** track personality dynamics where unsuccessful persuasion corresponds to lower levels of agreeableness and higher neuroticism throughout most of the conversation. The full corresponding conversations can be found in §H.

rski, 2023; Piskorski et al., 2023). Hence, we instead quantify performance by computing the Jensen–Shannon distance between the probe or prompt-derived $P(\text{strategy})$ distributions with that of a strong GPT-4 reference model (Lu et al., 2020; Ranathunga et al., 2024), as illustrated in Figure 7a. This approach is analogous to reference-based evaluation in NLP literature when ground-truth labels are absent (Lin, 2004; Liu et al., 2016).

Personality Trait Detection. We apply linear rescaling: $\text{score}_{1-5} = 1 + 4 \cdot P(\text{trait})$ to transform probability values $[0, 1]$ to the standard Big-5 scale $[1, 5]$ for direct comparison with human annotations. We then compute the mean squared error (MSE) for a given turn.

3 Identifying Features of Persuasion in Multi-Turn LLM Conversations

In this section, we compare the effectiveness of the techniques outlined in Section 2.3. We separately consider persuasion features at the sample-level (Sec. 3.1) and the population-level (Sec. 3.2). We broadly define persuasion features as text-based attributes that indicate or facilitate persuasive communication. This includes both positional markers which pinpoint where persuasive appeals occur in a dialogue, along with influential variables such as personality which causally contribute towards persuasion.

3.1 Can probes uncover persuasion features from individual samples?

Here, we manually examine a few examples in our datasets (chosen randomly) and show that probes can uncover various persuasive features in multi-turn dialogue.

Persuasion: Figure 4a shows the result of applying our coarse-grained persuasion probes to various samples in PfG, which captures distinct behavioral dynamics across various conversations (complete transcripts available in §H). In the unpersuasive sample, we observe a sharp drop in persuasion probability that directly corresponds to explicit rejection moments in the conversation ("I just can't see doing that", "I choose my own charities independently"). The trajectory also shows an upward spike on turns 19 and 20, coinciding with conversation-closing positive sentiment ("Have a nice day!"). This shows that our probe accurately detects the rejection moment but conflates later positive sentiment with persuasion. The trajectory for the persuaded sample appears smoother than the unpersuaded sample, which reflects EE-ER dynamics in the corresponding conversation. In the persuaded sample, the EE tends to agree more with the ER while in the unpersuaded sample, the EE asks more questions and is skeptical about the ER's arguments.

Personality: In Figures 4b and 4c, we visualize agreeableness and neuroticism trajectories

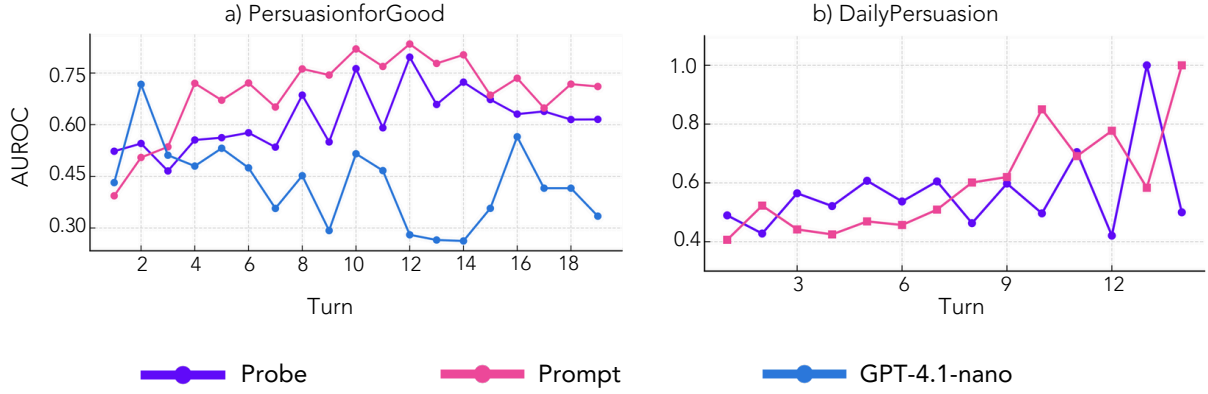


Figure 5: **Persuasion trajectory evaluation** across **a) PersuasionforGood** (OOD) and **b) DailyPersuasion** (synthetic). On our synthetic dataset, probes achieve perfect AUROC on the second-last conversation turn, while on OOD, AUROC peaks at mid-conversation. These performance maxima precisely align with the moments in the conversation where persuasion occurs. Low AUROC regions correspond to conversational segments dominated by non-indicative language patterns, such as greetings and formalities. Both probing and prompting the underlying model (pink line), have nearly identical performance maxima at roughly the same turn. See §H.4, H.5 for conversation transcripts from both datasets.

for a persuaded and an unpersuaded conversation (§H.3). For the unpersuaded conversation, we see consistently low agreeableness scores and high neuroticism scores throughout most of the conversation, which reflects the underlying tension and argumentative tone between the EE and ER. Further, the highest neuroticism regions occur at turns 2-5, which is the region where the EE expressed their unwillingness to donate. Conversely, the persuaded conversation yields a lower neuroticism throughout most turns and generally higher agreeableness, reflecting the more positive tone between participants. The low, 1.0 neuroticism score at turn 9 of the persuaded sample corresponds precisely to when the EE agreed to donate. There is a sharp drop in agreeableness at turn 6 corresponding to a clarifying question from the EE, suggesting the probe may misconstrue information-seeking behavior as resistance. *These patterns suggest potential correlations between Big-5 personality trait scores and persuasion, a connection we explore in more detail in Section 4.*

Strategy: Figure 4d shows a conversation dominated by credibility-based persuasion (§H.2). The highest probability peaks at turns 11, 17, and 19 directly correspond to explicit credibility appeals from the ER. However, before turn 10 there are no clear credibility appeals, despite having a moderately high credibility probability.

These findings demonstrate how **our behavioural probes can identify conversation segments with specific persuasive characteristics**. Probes

trained to predict persuasion strategy can capture when in a conversation a particular strategy such as ‘credibility appeal’ is used. Further, they can **reveal meaningful correlations between personality traits and persuasion outcomes**, for instance a higher neuroticism score correlating with being unpersuaded. This differs from recent probing literature, which has typically focused on detecting abstract behaviors like deception or political affiliation in isolation, without examining important sub-behaviors that constitute such phenomena (Kim et al., 2025; Goldowsky-Dill et al., 2025).

3.2 Do these features generalize across the entire dataset?

We previously saw that our behavioural probes can capture meaningful persuasion dynamics within individual conversation samples. We now extend our analysis to dataset-level behaviours, showing how we can uncover population-wide persuasion behaviour, a dimension that is lacking in recent work on creating persuasion datasets (Ma et al., 2025b; Jin et al., 2024; Zhang and Zhou, 2025).

Persuasion: Figure 5 shows persuasion classification AUROC across conversation turns for various models on both (a) OOD human-human and (b) synthetic LLM-LLM datasets. Examining the probe performance trajectory on the OOD dataset, performance plateaus during early (turns 1-7) and late (turns 16-19) conversation phases. This di-

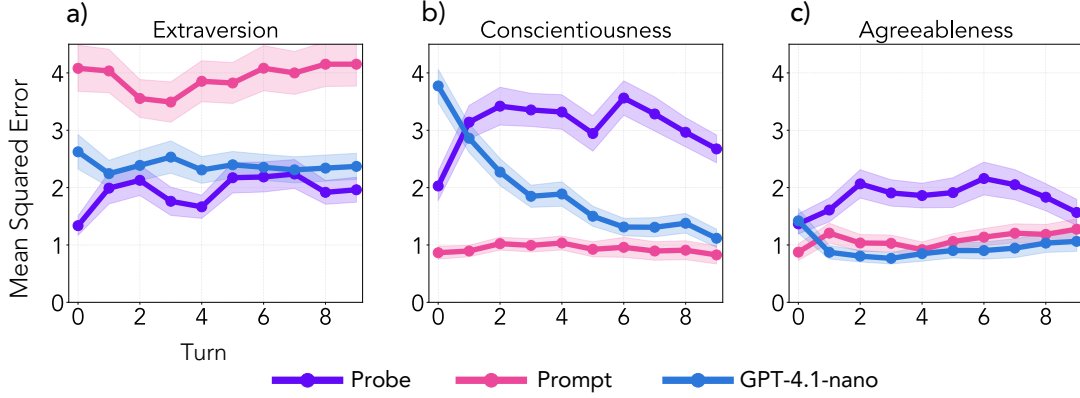


Figure 6: **Mean squared error (MSE) for predicting persuadee personality across various models.** Probes do slightly better than prompting and GPT-4.1-Nano for predicting **a)** extraversion but underperform on **b)** conscientiousness and **c)** agreeableness. *Ultimately, our results here are varied with no single model doing best at predicting personality traits.* However, considering model predictions in aggregate tells us which traits may be more difficult to predict from text. For instance, the models have a relatively lower MSE for agreeableness compared to extraversion suggesting that the former may be a more salient feature in the text. Plots for the remaining personality traits and *persuader* personality can be found in Figures 11 and 12 (§B).

rectly corresponds to low persuasion signal turns² dominated by formalities, greetings/goodbyes, and contextual framing where the ER establishes the donation setting. The highest classification performance occurs on EE turns 8, 10, and 12 – the conversation midpoint where EEs typically express their donation intentions. Random samples from PfG (see §H.4) corroborate these persuasion signal trends. The prompting trajectory follows a similar pattern, with performance peaks at the same middle turns and lower performance at the early and late turns. This shows, surprisingly, that added context *after* the EE makes their persuasion decision can hinder model performance at classifying persuasion. See for example §H.4 where the EE expresses willingness to donate on turn 10/20, followed by a discussion on conversation logistics and closing formalities. This subsequent context can mislead the model toward predicting unpersuasion, possibly because the ER provides additional reasons why the EE *should not* donate after they have already expressed their positive decision.

On DP we see a rather different persuasion trajectory with consistently low performance throughout most of the conversation, followed by a sharp performance spike during the final turns, consistent across both models. This pattern reflects DP’s structural characteristics (§H.5), where the EE is typically skeptical of the ER’s arguments through-

out most of the conversation, and makes their decision on the last 1-2 turns. This differs from PfG where the persuasion signal is in the middle of the conversation rather than toward the end.

Personality: In Figure 6, we compare various models’ ability to predict personality traits across turns. We focus on predicting persuadee personality since we want to measure how susceptible persuadees are to various persuasion strategies. *Our results here are varied with no single model outperforming the rest at predicting personality across turns.* These mixed results likely stem from the inherent difficulty of predicting personality from brief conversational text. For instance, it seems hard to deduce anything about the persuadee’s introversion after reading the PfG sample in §H.4 a).

For predicting conscientiousness, prompting consistently outperforms the other models while for extraversion it consistently underperforms. Probing consistently works best when predicting extraversion. A unique property of GPT-4.1-Nano is that its MSE roughly monotonically decreases with turn, indicating that it gets better at predicting personality as the conversation progresses. Interestingly, for agreeableness and neuroticism (§B) model MSEs are closer together compared to extraversion and conscientiousness. This might suggest that the former two traits are more salient within the text due to consistently low predictions across models.

Strategy: In Figure 7 we compare probes against prompting at uncovering persuasion strat-

²Turns that contain “persuasion signal” are ones that contain information that can be used to determine persuasive success.

egy, using GPT-4.1-Nano as a reference model. To achieve this, for each turn, we compute the Jensen-Shannon distance between the empirical distribution over strategies for probing/prompting against GPT-4.1-Nano (see Figure 10, §B for more details). From Figure 7b, across most turns, our probes share a more similar strategy distribution to GPT-4.1-Nano over prompting. Both probes and GPT-4.1-Nano show credibility as the dominant strategy across most conversation turns, in-line with the analysis from the PfG paper (Wang et al., 2019). In GPT-4.1-Nano, emotional appeal dominates in the early turns (1-3) which may be interpreted as labeling greetings as an emotional appeal since that is what these turns mostly consist of. Ultimately, these results indicate that probing works better than prompting the underlying model for predicting persuasion strategy trajectory.

The above represent *context-based* evaluations, where models analyze turns with accumulated prior context. We additionally conducted *context-less* evaluations in Table 3 (§F), finding that our persuasion probe achieves near perfect accuracy when classifying turns across both datasets. Moreover, in Figure 18 (§G.1), using PfG’s human-annotated semantic labels (agree/disagree donation, greeting, etc.), we found that our persuasion probe is “semantically calibrated” in that it assigns higher probabilities to text that is more indicative of positive persuasion.

Overall, our persuasion and strategy probes, can capture persuasive behaviour in the underlying datasets. The persuasion probe reveals how persuasion signal is distributed throughout a conversation, while the strategy probe identifies credibility as the dominant approach, consistent with GPT-4.1-Nano and Wang et al.’s (2019) assessments. With personality detection, we observe greater ambiguity, with no single model consistently outperforming others across all traits, but that using all models together as an ensemble can tell us which traits are more salient within the underlying text. These positive results demonstrate that **linear probes offer an efficient alternative for performing large-scale dataset analysis**, providing insights that would be too expensive to obtain through prompting-based methods alone.

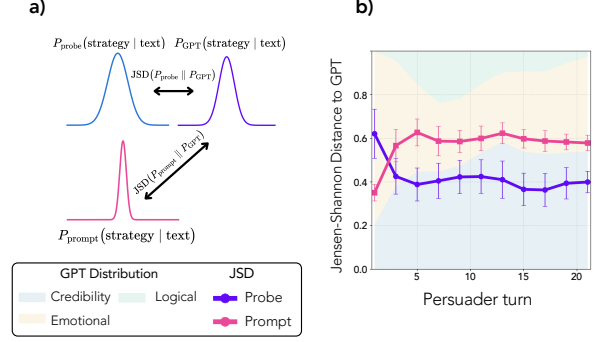


Figure 7: **Evaluating persuasion strategy across probing and prompting.** a) illustrates our approach where at each turn, we compute the Jensen-Shannon distance (JSD) between the mean of the probe’s and prompt’s strategy distributions and the GPT reference distribution. Our results are shown in b) where across most turns, the probe’s distribution more closely matches GPT than the prompt’s distribution, with both methods consistently highlighting credibility as the dominant strategy throughout the conversation. See Figure 10 (§B) for more details on this computation.

4 The Influence of Rhetorical Strategy and Personality on Persuasion

In this section, we apply our linear probes to identify and analyze correlations between persuader strategy and persuadee personality, exploring how these factors influence persuasion outcomes both in isolation and jointly.

4.1 Can we detect (un)-persuasion based on persuadee personality traits?

In Figure 4 we observed a few samples where a higher neuroticism and lower agreeableness score trajectories corresponded to unpersuasive conversations. We systematically validate whether this holds true across the entire dataset by applying the corresponding behavioral probes to the full dataset as shown in Figure 8. We flag any sample with $P(\text{agreeableness}) < 0.2$ or $P(\text{neuroticism}) > 0.8$ as unpersuasive, and find that turns 3–5 recover a large fraction of unpersuaded cases while keeping persuasion false alarms low. *Notably, this same turn window matches the high-accuracy region in the persuasion AUROC curve (Figure 5a), reinforcing a potential link between these personality trait values and unpersuasion.* There remain some missed unpersuasion cases and falsely flagged persuasion examples, since low agreeableness and high neuroticism are not strict conditions for unpersuasion. An EE that scores low in agreeableness can still choose to donate (see false persuasion

alarms in §H.6), and an EE that scores high in agreeableness can still decide not to donate (see missed unpersuasion cases in §H.7). Interestingly, however, the inverse is not true: high agreeableness and low neuroticism scores do not pick out *persuasive* samples well (Figure 13 §B).

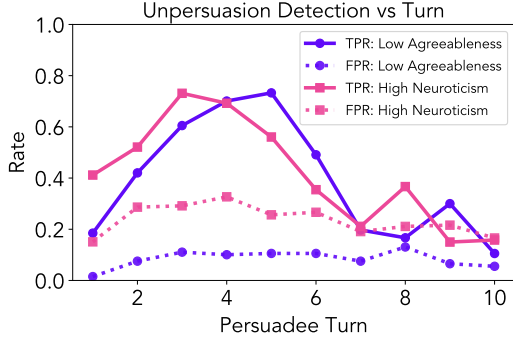


Figure 8: **Unpersuasion detection vs turn** using critical personality trait regions discovered in Figure 4. Treating unpersuasion as the positive class, we evaluate whether low agreeableness (thresh < 0.2) and high neuroticism (thresh > 0.8) can reliably flag unpersuaded cases (TPR) while keeping persuasion false alarms (FPR) low. Indeed, we see this can be achieved on turns 3-5, corresponding to roughly the same high persuasion accuracy regions in Figure 5a.

4.2 What correlations exist between rhetorical strategy and personality?

In Figure 9, we compute correlations between persuasion strategies and personality traits across both PfG and DP datasets, using scores produced by our probes. Across both datasets, extroversion is moderately correlated with various persuasion strategies, which is surprising since these datasets are quite different. In PfG, extroversion shows a moderately positive correlation with emotional appeal while having a negative correlation with credibility appeal. The former correlation was also found in the original PfG paper (Wang et al., 2019). This pattern is partially replicated in DP, where extroversion again correlates positively with emotional appeal but now shows a negative correlation with logical appeal. These consistent findings across datasets may suggest that extroversion can be a key mediator in persuasion dynamics where extraverted individuals are more susceptible to emotional appeal while being less susceptible to logical and credibility appeals. They may also *hint at the lack of diversity* in existing datasets, since we might expect more correlations, such as agreeableness being more susceptible to persuasion strategies in com-

parison to openness (Alkış and Taşkaya Temizel, 2015).

Note that while the individual correlation values are not that high, their relative magnitude compared to other trait-strategy correlations and their consistency across disparate datasets makes them noteworthy. Finally, these findings should be taken lightly due to the difficulty in predicting extraversion from text, as evidenced by the high model disagreement shown in Figure H.3 (§B).

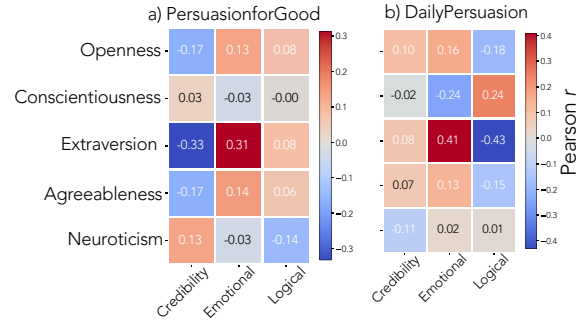


Figure 9: **Correlations between EE personality and ER strategy.** Consistent with findings in PfG, our probes find a moderate positive correlation between extroversion and emotional appeal. However, we also identify a moderate negative correlation between extroversion and credibility, not previously discovered. In DP, we novelly find a moderate-strong negative correlation between extroversion and logical appeal along with a moderate-strong positive correlation between extroversion and emotional appeal, similar to in PfG.

5 Discussion

In this work, we applied linear probes to understand how LLMs persuade in multi-turn conversations. We used insights from cognitive science to probe LLMs for persuasion and its various behavioral components: persuasion strategy and persuadee personality. Despite their simplicity, we found probes work surprisingly well at revealing meaningful features of persuasive conversations such as the point in a conversation where the persuadee was unpersuaded and the interaction effects of strategy and personality on persuasion. We found that, in the synthetic dataset, persuasive success is confined to the last one or two turns, whereas in the human dataset it peaks around the conversation midpoint. The successful application of probes in our work, suggests a promising avenue for future work to apply probes to understanding other abstract behaviours in LLMs such as deception and manipulation. This is especially important in multi-turn set-

things and when conducting large-scale dataset analysis, where prompting-based approaches would be computationally inefficient.

6 Limitations and Future Work

We focus on credibility, logical, and emotional appeals as a minimal basis into which all other strategies can be categorized. More granular strategies like foot-in-the-door or task-related inquiry (Wang et al., 2019) could yield additional insights. Similarly, while we use the Big-5 personality framework, alternative models such as the Moral Foundations questionnaire (Graham et al., 2011) can give us additional insights into how morality can influence persuasion. Beyond personality and strategy, our work can extend to other persuasion-relevant behaviors, such as tone of voice (Vaughan-Johnston et al., 2021) or persuader-persuadee relationship dynamics (Markowitz, 2025) to see how these contribute towards persuasion. It also would be interesting to see how our behavioral probes can be used to detect harmful persuasive behavior in high-stakes domains like insider trading (Scheurer et al., 2024).

In our experiments, we extract activations from Llama-3.2-3b; testing larger variants and different model architectures would strengthen generalization claims. However, our promising results with a 3b model potentially suggest scalability to larger models.

7 Potential Risks

The risks resulting from this work are minimal as we are not creating models to generate persuasive text nor proposing new persuasion datasets or benchmarks. However, our probes reveal interaction effects between persuasion strategies and personality traits that could potentially be misused to develop targeted persuasion techniques for specific psychological profiles. Further, our evaluation may amplify potential biases present in the underlying datasets. For PfG, this includes potential demographic biases of human participants while for DP this can be biases in the GPT-generated synthetic data generation framework. Nevertheless, these are still valuable insights providing new perspective on the persuasion dynamics of the underlying datasets.

Acknowledgements

We thank Elliot Creager and David Duvenaud for helpful feedback and discussions that contributed

towards this draft. This research was funded through a grant from the Berkeley Existential Risk Initiative (BERI) along with computing infrastructure provided by Mila (mila.quebec).

References

- Guillaume Alain and Yoshua Bengio. 2018. [Understanding intermediate layers using linear classifier probes](#). *Preprint*, arXiv:1610.01644.
- Nurcan Alkış and Tuğba Taşkaya Temizel. 2015. [The impact of individual differences on influence strategies](#). *Personality and Individual Differences*, 87:147–152.
- Aristotle. 1932. *Rhetoric*. Loeb Classical Library. Harvard University Press, Cambridge, MA. Original work published ca. 350 BCE; Book I, Chapters 2–3.
- Hui Bai and 1 others. 2025. [Llm-generated messages can persuade humans on policy issues](#). Web. Preprint.
- Pablo Briñol and Richard E. Petty and. 2009. [Source factors in persuasion: A self-validation approach](#). *European Review of Social Psychology*, 20(1):49–96.
- Carlos Carrasco-Farre. 2024. [Large language models are as persuasive as humans, but how? about the cognitive effort and moral-emotional language of llm arguments](#). *Preprint*, arXiv:2404.09329.
- Shelly Chaiken. 1980. [Heuristic versus systematic information processing and the use of source versus message cues in persuasion](#). *Journal of Personality and Social Psychology*, 39(5):752–766.
- Canyu Chen and Kai Shu. 2024. [Combating misinformation in the age of llms: Opportunities and challenges](#). *AI Mag.*, 45(3):354–368.
- Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. 2024. [Cogbench: a large language model walks into a psychology lab](#). In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. [Measuring the persuasiveness of language models](#).
- Alice H. Eagly and Shelly Chaiken. 1984. [Cognitive theories of persuasion](#). *Advances in Experimental Social Psychology*, 17:267–359.
- Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. 2025. [Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars](#). *Preprint*, arXiv:2503.01307.

- S. García-Méndez, F. de Arriba-Pérez, and M. d. C. Somoza-López. 2025. [A review on the use of large language models as virtual tutors](#). *Science & Education*, 34:877–892.
- Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. 2025. [Detecting strategic deception using linear probes](#). *Preprint*, arXiv:2502.03407.
- Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. [How persuasive is ai-generated propaganda?](#) *PNAS Nexus*, 3(2):pgae034.
- Jesse Graham, Brian A. Nosek, Jonathan Haidt, Rachna Iyer, Slavica Koleva, and Peter H. Ditto. 2011. [Mapping the moral domain](#). *Journal of Personality and Social Psychology*, 101(2):366–385.
- Anthony G. Greenwald. 1968. Cognitive learning, cognitive response to persuasion, and attitude change. In Anthony G. Greenwald, Timothy C. Brock, and Thomas M. Ostrom, editors, *Psychological Foundations of Attitudes*, pages 147–170. Academic Press, New York.
- Wes Gurnee and Max Tegmark. 2024. [Language models represent space and time](#). In *The Twelfth International Conference on Learning Representations*.
- Kobi Hackenburg and Helen Margetts. 2024. [Evaluating the persuasive influence of political microtargeting with large language models](#). *Proceedings of the National Academy of Sciences*, 121(24):e2403116121.
- Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. [Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt](#). *Nature Computational Science*, 3(10):833–838.
- Hans Hoeken and Daniel J. O’Keefe. 2022. [The reconstructability of persuasive message variables affects the variability of experimental effect sizes: Evidence and implications](#). *Human Communication Research*, 48(4):543–552.
- Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. [Persuading across diverse domains: a dataset and persuasion large language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.
- Oliver P. John, Eileen M. Donahue, and Robert L. Ken-
tle. 1991. Big five inventory (bfi). Database record, APA PsycTests.
- Harold H. Kelley and John L. Michela. 1980. [Attribution theory and research](#). *Annual Review of Psychology*, 31:457–501.
- Junsol Kim, James Evans, and Aaron Schein. 2025. [Linear representations of political perspective emerge in large language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Jinghui Lu, Maeve Henchion, and Brian Mac Namee. 2020. [Diverging divergences: Examining variants of Jensen Shannon divergence for corpus comparison tasks](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6740–6744, Marseille, France. European Language Resources Association.
- Floridi Luciano. 2024. [Hypersuasion – on AI’s persuasive power and how to deal with it](#). *Philosophy & Technology*, 37:64.
- Weicheng Ma, Hefan Zhang, Ivory Yang, Shiyu Ji, Joice Chen, Farnoosh Hashemi, Shubham Mohole, Ethan Gearey, Michael Macy, Saeed Hassanpour, and Soroush Vosoughi. 2025a. [Communication is all you need: Persuasion dataset construction via multi-llm communication](#). *Preprint*, arXiv:2502.08896.
- Weicheng Ma, Hefan Zhang, Ivory Yang, Shiyu Ji, Joice Chen, Farnoosh Hashemi, Shubham Mohole, Ethan Gearey, Michael Macy, Saeed Hassanpour, and Soroush Vosoughi. 2025b. [Communication makes perfect: Persuasion dataset construction via multi-LLM communication](#). 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 4017–4045, Albuquerque, New Mexico. Association for Computational Linguistics.
- D. M. Markowitz. 2025. [Language style \(mis\)matching: Consuming entertainment media from someone unlike you is linked to positive attitudes](#). *Scientific Reports*, 15:1894.
- Samuel Marks and Max Tegmark. 2024. [The geometry of truth: Emergent linear structure in large language model representations of true/false datasets](#). In *First Conference on Language Modeling*.
- W. J. McGuire. 1972. Social psychology. In P. C. Dodwell, editor, *New horizons in psychology*. Penguin.
- Alex McKenzie, Urja Pawar, Phil Blandfort, William Banks, David Krueger, Ekdeep Singh Lubana, and

- Dmitrii Krashennnikov. 2025. [Detecting high-stakes interactions with activation probes](#). *Preprint*, arXiv:2506.10805.
- Meta. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. [Meta AI blog post on Llama 3.2](#). Accessed: 2025-05-17.
- NVIDIA Corporation. 2020. NVIDIA A100 Tensor Core GPU Architecture. <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf>. Accessed: 2025-05-17.
- Nick Obradovich, Sahib S. Khalsa, Wasiq Ullah Khan, Jessica Fitzpatrick, and Kimberly Young. 2024. [Opportunities and risks of large language models in psychiatry](#). *NPP—Digital Psychiatry and Neuroscience*, 2:8.
- OpenAI. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2025. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>. Accessed: 2025-05-17.
- Kiemute Oyibo and Julita Vassileva. 2019. [The relationship between personality traits and susceptibility to social influence](#). *Computers in Human Behavior*, 98:174–188.
- Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. [Measuring and benchmarking large language models’ capabilities to generate persuasive language](#). *Preprint*, arXiv:2406.17753.
- Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023. [Multilingual multifaceted understanding of on-line news in terms of genre, framing, and persuasion techniques](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.
- Surangika Ranathunga, Nisansa De Silva, Velayuthan Menan, Aloka Fernando, and Charitha Rathnayake. 2024. [Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 860–880, St. Julian’s, Malta. Association for Computational Linguistics.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. [On the conversational persuasiveness of large language models: A randomized controlled trial](#). *ArXiv*, abs/2403.14380.
- Jérémy Scheurer, Mikita Balesni, and Marius Hobbahn. 2024. [Large language models can strategically deceive their users when put under pressure](#). In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.
- Rusheb Shah, Quentin Feuillade-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. 2023. [Scalable and transferable black-box jail-breaks for language models via persona modulation](#). *Preprint*, arXiv:2311.03348.
- L. Shen and E. Bigsby. 2013. The effects of message features: Content, structure, and style. In J. P. Dillard and L. Shen, editors, *The SAGE handbook of persuasion: Developments in theory and practice*, 2nd edition, pages 20–35. Sage Publications, Inc.
- Somesh Kumar Singh, Yaman Kumar Singla, Harini S I, and Balaji Krishnamurthy. 2025. [Measuring and improving persuasiveness of large language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Emily C. Stadel, Shannon Wiltsey Stirman, Lyle H. Ungar, and Johannes C. Eichstaedt. 2024. [Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation](#). *npj Mental Health Research*, 3:12.
- Nicolas Stefanovitch and Jakub Piskorski. 2023. [Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 71–86, Singapore. Association for Computational Linguistics.
- M. Stella, T. T. Hills, and Y. N. Kenett. 2023. [Using cognitive psychology to understand gpt-like models needs to extend beyond human biases](#). *Proceedings of the National Academy of Sciences of the United States of America*, 120(43):e2312911120. Epub 2023 Oct 16; PMID: 37844246; PMCID: PMC10614207.
- Brian Sternthal, Lynn W. Phillips, and Ruby Dholakia. 1978. [The persuasive effect of source credibility: A situational analysis](#). *The Public Opinion Quarterly*, 42(3):285–314.
- Curt Tigges, Oskar John Hollinsworth, Neel Nanda, and Atticus Geiger. 2024. [Language models linearly represent sentiment](#).
- Zakary L. Tormala, Pablo Briñol, and Richard E. Petty. 2006. [When credibility attacks: The reverse impact of source credibility on persuasion](#). *Journal of Experimental Social Psychology*, 42(5):684–691.

- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2025. [Steering language models with activation engineering](#).
- T. I. Vaughan-Johnston, J. J. Guyer, L. R. Fabrigar, and Others. 2021. [The role of vocal affect in persuasion: The civa model](#). *Journal of Nonverbal Behavior*, 45:455–477.
- Xuwei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. [Persuasion for good: Towards a personalized persuasive dialogue system for social good](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.
- Zhenhua Wang, Wei Xie, Baosheng Wang, Enze Wang, Zhiwen Gui, Shuoyoucheng Ma, and Kai Chen. 2024. [Foot in the door: Understanding large language model jailbreaking via cognitive psychology](#). *Preprint*, arXiv:2402.15690.
- Nan Xu, Fei Wang, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. 2024. [Cognitive overload: Jailbreaking large language models with overloaded logical thinking](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3526–3548, Mexico City, Mexico. Association for Computational Linguistics.
- Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. [How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350, Bangkok, Thailand. Association for Computational Linguistics.
- Dingyi Zhang and Deyu Zhou. 2025. [Persuasion should be double-blind: A multi-domain dialogue dataset with faithfulness based on causal theory of mind](#). *ArXiv*, abs/2502.21297.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. [Representation engineering: A top-down approach to ai transparency](#). *Preprint*, arXiv:2310.01405.

Appendix

Table of Contents

A	Related Work	15
B	Main Paper Figure Extensions	16
C	Sample Sizes and Statistical Significance	17
C.1	Upsampling to increase the amount of unpersuasive samples	17
D	Probe Learning Curves	18
E	Data Analysis and Validation	20
E.1	Validating where persuasion occurs in a conversation	20
E.2	DailyPersuasion (DP) Data Analysis	20
F	Evaluating Probes on Individual Utterances	23
G	Persuasion Probe Interpretability	24
G.1	Experiments	24
H	Conversation Transcripts	26
H.1	Persuasion Probability Samples for Figure 4 a)	26
H.2	Credibility Appeal Sample for Figure 4 b)	27
H.3	Personality Samples for Figure 4 c) and d)	28
H.4	Conversation Samples from PersuasionforGood	29
H.5	Conversation Samples from DailyPersuasion	30
H.6	Agreeableness False Alarms for Section 4.1	31
H.7	Agreeableness Missed Unpersuasion for Section 4.1	33
I	Synthetic Data Generation for Training Probes Details	35
I.1	Prompt Template	35
I.2	Conversation Transcripts	35
I.3	Sample Contexts, Scenarios, Roles, and Archetypes	37
I.4	Synthetic Probe Training Dataset Distributional Statistics	38
J	Responsible NLP Checklist	39
J.1	Licenses Used for Data and Models	39
J.2	More Experimental Details	39
J.3	Artifact Documentation and Data Statistics	39
J.4	Annotation Procedures	40

A Related Work

Linear Probing: Linear probing was classically applied by [Alain and Bengio \(2018\)](#) to understand linear separability in intermediate layers of computer vision models. Recent works have extended this technique to understand how LLMs represent abstract phenomena such as space-time ([Gurnee and Tegmark, 2024](#)), political perspective ([Kim et al., 2025](#)), and truth ([Marks and Tegmark, 2024](#)). Our work represents the first application of probing to persuasion and its constituent behaviors like rhetorical strategy and personality traits. Unlike prior efforts that typically analyze individual behaviors in isolation, we are the first to demonstrate how probing multiple constituent behaviors can provide insights into a single complex cognitive phenomena—persuasion. Further, the probing literature focuses on the single-turn setting with [Goldowsky-Dill et al. \(2025\)](#) being the only other work we are aware of that focuses on multi-turn in the context of uncovering strategic deception. Lastly, alternative approaches for analyzing model internals include unsupervised PCA ([Zou et al., 2025](#)) and activation difference vectors ([Turner et al., 2025](#); [Rimsky et al., 2024](#)), but we opt for linear probes since they have been more widely adopted with great success.

Persuasion in Cognitive Science: Cognitive science and psychology offer a rich theoretical foundation for understanding persuasion dynamics which can offer additional insights into our empirical results. Receiver-focused theories frame persuasion as primarily occurring within the message recipient ([Eagly and Chaiken, 1984](#)), including dual-process frameworks like the heuristic-systematic model (HSM) ([Chaiken, 1980](#); [Greenwald, 1968](#)), causal attribution theory ([Kelley and Michela, 1980](#)), and probabilistic models ([McGuire, 1972](#)). Alternative perspectives emphasize the sender’s role ([Briñol and and, 2009](#); [Sternthal et al., 1978](#); [Tormala et al., 2006](#)) or message content characteristics ([Shen and Bigsby, 2013](#); [Hoeken and O’Keefe, 2022](#)). Each of these perspectives can be adopted to further explain the empirical results in our work. For example, in Figure 4c and 4d, where an EE with low agreeableness and high neuroticism remains unpersuaded, HSM might suggest the EE uses central-route processing—critically evaluating arguments through self-dialogue rather than responding to peripheral cues like emotional appeals or source credibility. Conversely, a message-content perspective might attribute the failed persuasion attempt to structural weaknesses in the message itself such as poor grammar or increased verbosity. Future work should investigate this further and more generally use these theoretical frameworks to explain persuasive behaviors between AI agents and humans.

AI-Based Persuasion: Recent work has focused on optimizing AI to generate persuasive dialogue ([Jin et al., 2024](#); [Zhang and Zhou, 2025](#); [Ma et al., 2025a](#)), evaluating persuasion effectiveness ([Durmus et al., 2024](#); [Pauli et al., 2025](#); [Singh et al., 2025](#)), and on behavioral studies investigating the persuasive impact of AI on humans ([Luciano, 2024](#); [Salvi et al., 2024](#); [Carrasco-Farre, 2024](#)). Most existing persuasion datasets used in these works lack the dynamic, multi-turn analysis of persuasion that our approach provides, making our work a valuable complement to persuasion dataset development. Recent work on jailbreaking is also a form of persuasion as humans must persuade LLMs to override their safety guardrails ([Zeng et al., 2024](#); [Xu et al., 2024](#); [Shah et al., 2023](#)). In contrast, our research focuses on the understudied scenario of LLMs as persuaders in naturalistic, multi-turn interactions with human or LLM persuadees.

Cognitive Science Applied to LLMs: Ideas from cognitive science have recently informed LLM research across diverse domains such as reasoning ([Gandhi et al., 2025](#); [Hagendorff et al., 2023](#)), jailbreaking ([Wang et al., 2024](#); [Zeng et al., 2024](#)), and evaluating cognitive ability ([Coda-Forno et al., 2024](#); [Stella et al., 2023](#)). In our work, we use insights from cognitive science to guide our synthetic data generation process by simulating persuader-persuadee interactions conditioned on cognitive dimensions such as personality traits and rhetorical strategies. They also inform our probe design by leveraging established empirical findings on the relationships between persuadee personality, persuasion strategy, and persuasion outcomes ([Oyibo and Vassileva, 2019](#); [Alkış and Taşkaya Temizel, 2015](#)).

B Main Paper Figure Extensions

In this section, we provide extensions to Figures 6, 7, and 8 in the main paper. Please see the corresponding figure descriptions below for a summary of the extended result.

- Figure 10 is the extension to Figure 7 which shows the full strategy trajectory curves for each model. Figure 7 is computed from this by taking a point-wise JSD between probe/prompt model and GPT-4.1-Nano, for a given turn.
- Figures 11 and 12 are full personality trajectory plots for each trait, extending Figure 6.
- Figure 13 extends Figure 8 and provides the negative result of the inverse problem discussed in Section 4.1.

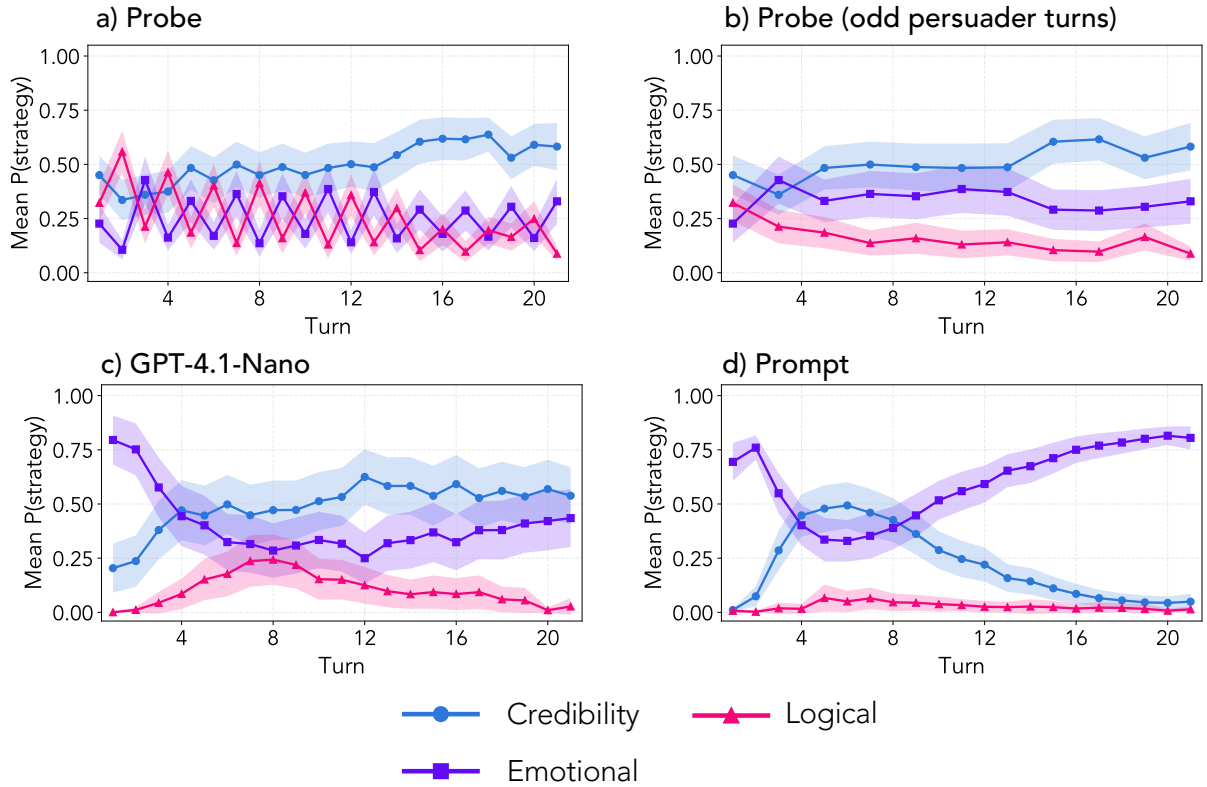


Figure 10: **Population-Level Strategy Trajectory Curves.** At each turn, we plot the mean of the empirical strategy distribution. Probing on odd persuader turns has a relatively identical strategy curve compared to GPT-4.1-Nano, the reference curve we’re comparing against. All plots show logical appeal as not dominant across all samples. The probe has competing ER and EE oscillations which smoothens out if we only consider even/odd turns. Note that as conversation length increases beyond 23 turns (a small subset of the total dataset), emotional appeal is the dominant strategy which is consistent across all models. A single turn corresponds to a single EE/ER utterance. In Figure 7b, we take a point-wise Jensen-Shannon Distance at each turn using the mean probability vectors in this plot.

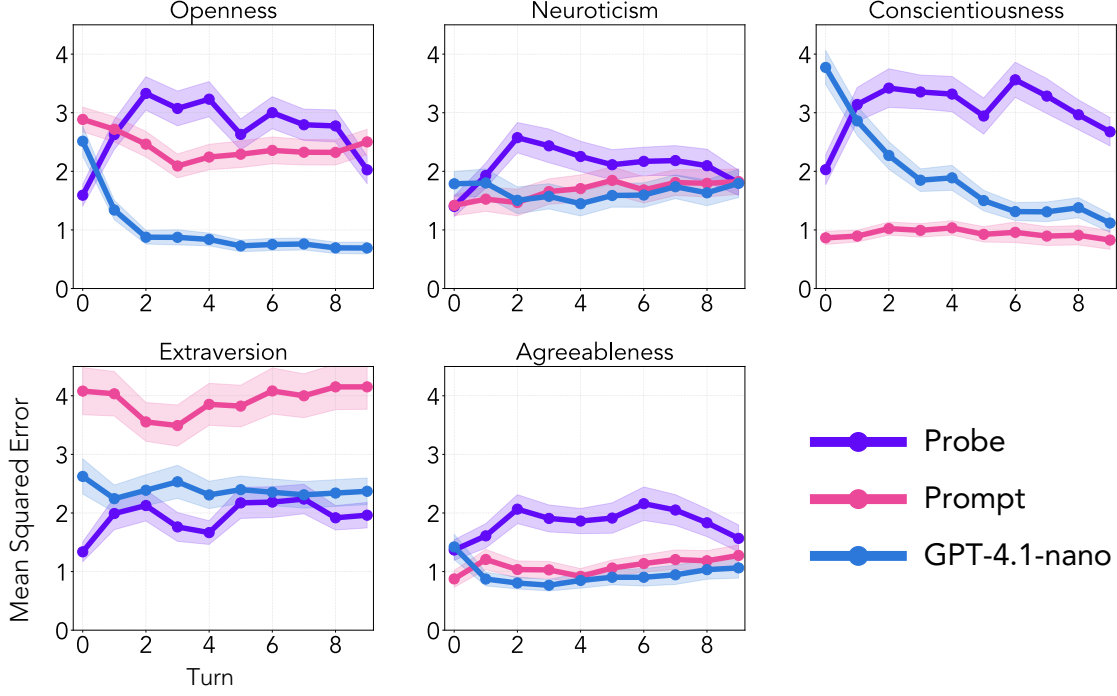


Figure 11: **Persuadee Personality Predictions**

C Sample Sizes and Statistical Significance

Table 1 outlines sample sizes and statistical significance for figures and experiments used in this paper.

Table 1: Samples Sizes used in Figures

Figure	Sample Size
Figure 4	Random samples.
Figure 5	PfG: $n = 401$ (232 persuasion, 169 unpersuasion); DP: $n = 186$ (93 per class, due to downsampling).
Figure 3	Timing comparison: synthetic token-level ($n = 5$), turn-level ($n = 10$), conversation-level ($n = 20$). Error bars represent average across 3 runs.
Figure 9	DP: $n = 557$, PfG: $n = 545$ (all positive donation samples)
Figures 10, 7	PfG: $n = 200$ for Probe/Llama-3b, $n = 140$ for GPT-4.1-Nano. Error bands represent 0.3σ .
Figures 6, 11	PfG: $n = 100$. Error bands represent 0.1σ .
Figure 12	PfG: $n = 20$ for Probe/Prompt, $n = 40$ for GPT. Error bands represent 0.1σ .
Figures 8, 13	PfG: $n = 401$ (since we filter to use annotated subset)

C.1 Upsampling to increase the amount of unpersuasive samples

As mentioned in Section 2.2, our evaluation datasets consist of predominantly positive samples. To remedy this, we few-shot prompt GPT-4.1 to increase the amount of unpersuasive samples. Table 2 shows the prompt template we use to upsample PfG unpersuasive samples (we do something similar for DP).

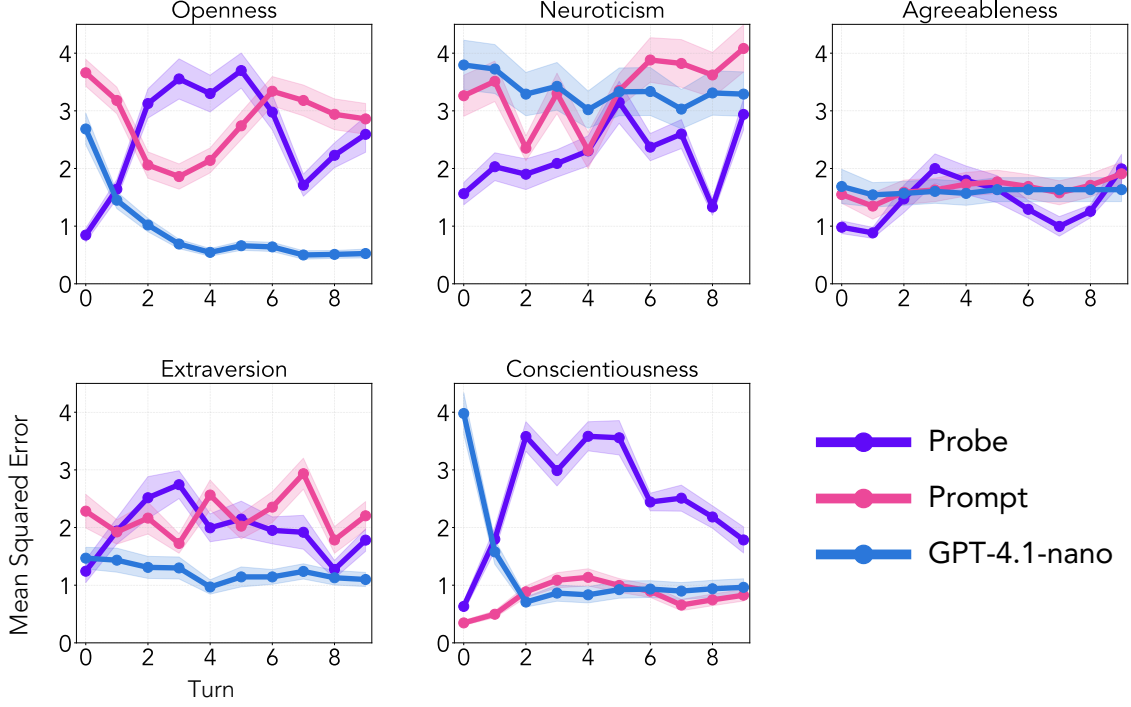


Figure 12: **Persuader Personality Predictions**

Table 2: Prompt Template for Unpersuasive Donation Conversations

System: You are a helpful assistant that writes realistic, multi-turn conversations in which a persuader fails to convince the persuadee to donate.

User: Below are N examples of conversations where the persuader fails to persuade the persuadee to donate (unpersuasive conversations):

Example 1: <Unpersuasive example 1 (omitted for brevity)>

Example 2: <Unpersuasive example 2 (omitted for brevity)>

...

Please generate k independent conversations in the same style. Each should be multi-turn, realistic, and unsuccessful at persuading the persuadee to donate.

D Probe Learning Curves

In Figure 14 we outline and show learning curves for various other linear probe configurations, aside from the context-based probe we use in the main paper.

These probes were all trained using the Adam optimizer with a learning rate of 1×10^{-3} , ReLU activation function, and the standard cross-entropy loss.

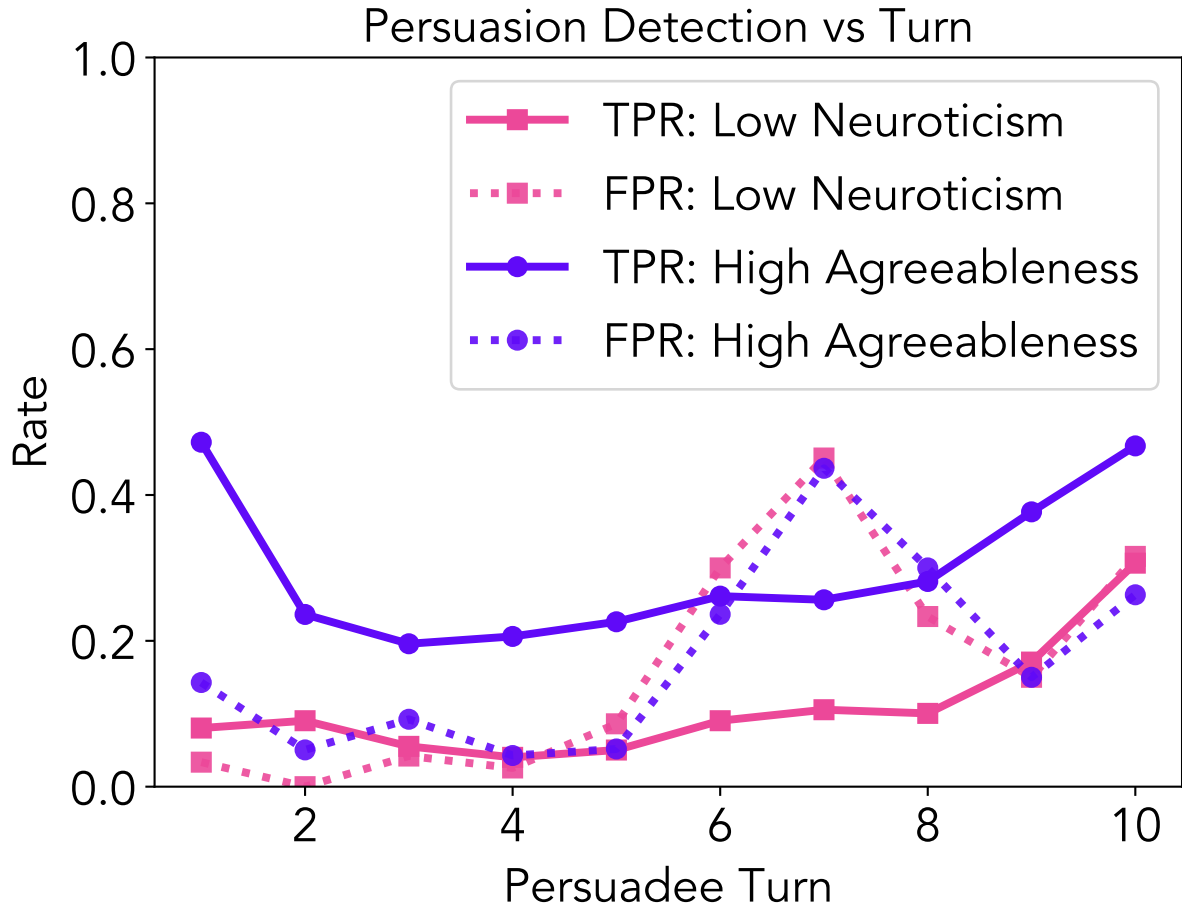


Figure 13: **Persuasion detection vs turn** represents the inverse of Figure 8: we test whether low neuroticism (thresh < 0.2) and high agreeableness (thresh > 0.8) can accurately identify persuasive dialogues (TPR) while minimizing unpersuaded false alarms (FPR). Neuroticism fails—the FPR curve consistently exceeds TPR, misflagging more unpersuaded than persuaded samples. Agreeableness does better with modest TPR peaks at the first and last turns, though not great overall.

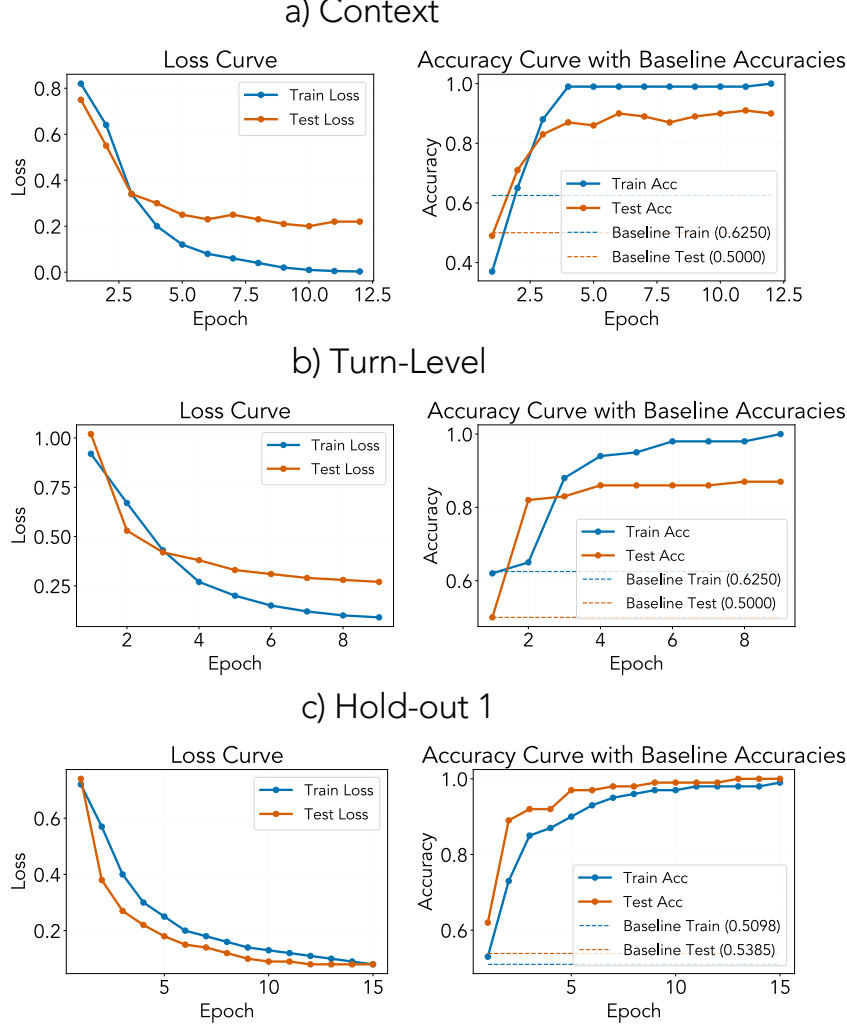


Figure 14: **Probe training curves across various settings.** The context probe a) is what we use in the main body of the paper. Additionally, we train (b) turn-level (contextless) probes which are trained on individual turns rather than entire conversations and (c) hold-out-1 probes which are trained on the entire conversation with the exception of the last turn. The idea behind (c) was that we observed most of the persuasion signal coming from the final conversation turn so removing this final turn could help mitigate a potential spurious correlation. However, probes a) and b) did not latch onto this spurious correlation and thus c) ended up under-performing compared to them.

E Data Analysis and Validation

E.1 Validating where persuasion occurs in a conversation

At several places in our paper, such as in Figure 5 we claim that in DailyPersuasion (DP), persuasion outcomes typically occur in the final 1-2 persuadee turns of a conversation, while in PersuasionforGood (PfG), they are more evenly dispersed throughout the middle-to-late turns. Here, we systematically test these claims by zero-shot prompting GPT-4.1 using the prompt template in Figure 16 to classify the turn where persuasion occurs. We use 56 random samples from DP and 100 from PfG. The first-author manually validated roughly 25 classifications with no disagreements on DP and 4/26 disagreements on PfG, indicating an overall high agreement. The resulting persuasion turn histograms are shown in Figure 15 which corroborate our claims.

E.2 DailyPersuasion (DP) Data Analysis

Here we show distributional statistics and an annotated conversation for DP (Jin et al., 2024).

Figure 17 shows distributions over both turn-level and conversation-level annotations generated by

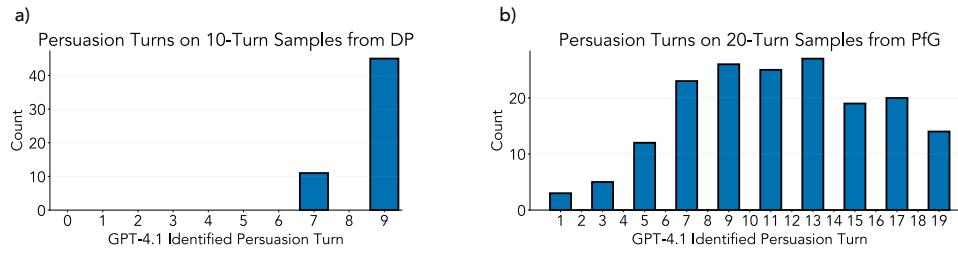


Figure 15: **Persuasion Turn Distributions** for **a)** DailyPersuasion (DP) and **b)** PersuasionforGood (PfG). These histograms show the frequency of successful persuasion moments by conversation turn. In DP, persuasion predominantly occurs in the final turn, while PfG exhibits a more even dispersion across the middle turns.

System:

Analyze when the persuadee was persuaded in this conversation.

Look for the specific turn where the persuadee’s attitude shifted or they became convinced.

User:

Turn 0 (Role): <first utterance>

Turn 1 (Role): <second utterance>

...

Turn N (Role): <Nth utterance>

Which turn number shows the persuadee being persuaded?

Reply with just the turn number (e.g. “3”) or “None” if no persuasion occurred.

Figure 16: System and user messages used to prompt GPT-4.1 for persuasion turn identification.

GPT-4o. We manually review these and find high human annotator agreement (Cohen’s $\kappa > 0.95$). See Section E.2 for a sample annotation.

Fully Annotated Success Conversation from DP

Fully Annotated Conversation, Conversation ID: 20230831164028-67_95_2

Session-Level Annotation (gpt-4o): success

Turn-Level Annotations (only for Persuadee turns):

Turn 1: not_persuadee
 Turn 2: non-persuasive
 Turn 3: not_persuadee
 Turn 4: uncertain
 Turn 5: not_persuadee
 Turn 6: non-persuasive
 Turn 7: not_persuadee
 Turn 8: uncertain
 Turn 9: not_persuadee
 Turn 10: uncertain
 Turn 11: not_persuadee
 Turn 12: persuasive

Dialogue Session:

Persuader: Hey Sam, did you know that volunteering at the local food bank can actually improve your problem-solving and teamwork skills? It's just like leveling up in a video game!

Persuadee: Really? But I still think I'd rather spend my weekends playing games. I don't see how volunteering can be that beneficial.

Persuader: You know, our friend Jessie started volunteering at the food bank last year. She told me that it helped her develop leadership skills, and she even got a scholarship for her community involvement! It's worth giving it a try, don't you think?

Persuadee: Hmm, I didn't know that. But what if I don't enjoy it? I'm not really a people person, and it sounds like a lot of work.

Persuader: I totally get that, but the great thing about volunteering at the

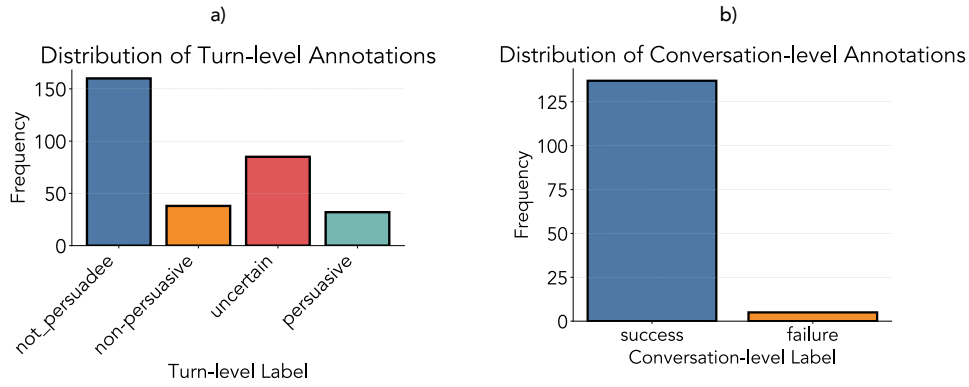


Figure 17: **Using GPT-4o to annotate DailyPersuasion turns and conversations.** Most conversations have positive persuasion motivating our choice to use GPT to up-sample failure conversations in Section 2.2. The persuasive/non-persuasive samples in turn-level annotations are used for the contextless evaluations in Table 3 (§F).

food bank is that we'll be working together as a team. Plus, you'll meet other people who share your interests, and it's always more fun to do things with friends, right?

Persuadee: Yeah, maybe. But I still think I'd rather just play video games. I don't feel like I have much to offer.

Persuader: Sam, you have so much to offer! Your gaming skills show that you're quick-thinking and resourceful, which are valuable qualities in a volunteer. Plus, you'll be making a real difference in people's lives by helping out at the food bank.

Persuadee: Well, maybe you're right. But I still have concerns about the time commitment. I don't want it to interfere with my gaming time.

Persuader: No worries, Sam! The food bank has flexible volunteering schedules, so you can choose a time that works best for you. You can still enjoy your gaming time and make a difference in the community.

Persuadee: Okay, that sounds more doable. Maybe I can give it a try, but I'm not making any promises.

Persuader: That's totally fine, Sam! Why not give it a try for just one or two weekends? If you don't like it, you can always stop, but at least you'll have given it a shot. I'll be there with you, and we can make it fun together!

Persuadee: Alright, I'll give it a try. But only for a couple of weekends, and then we'll see how it goes.

F Evaluating Probes on Individual Utterances

In this section, we evaluate how accurately various probe configurations (outlined in Figure 14, §D) can categorize individual persuadee utterances as “persuaded” or “not-persuaded” without access to the broader conversation context. We focus on the DailyPersuasion (DP) dataset, using the label distribution shown in Figure 17. For comparison, we also assess performance on (1) the full conversation, and (2) the full conversation with the final turn removed—since, as Figure 5 indicates, much of the persuasion signal in DP is concentrated in that last turn, potentially creating a spurious correlation to classifying persuasion.

Table 3: Comparison of probe performance on the DailyPersuasion dataset. Ultimately, the context probe does best, motivating our use of it in the main experiments of this paper.

Probe Configuration	Turn Classification			Conversation Classification					
	Accuracy	Precision	Recall	Full conversation			Last turn omitted		
				Accuracy	Precision	Recall	Accuracy	Precision	Recall
Context*	91	91	91	95	96	99	95	96	99
No-context	68	60	97	97	97	100	89	97	92
Hold-1	83	86	75	96	96	99	95	96	99
Hold-2	71	80	50	94	97	97	76	98	77

Results: Table 3 shows the result of our probe evaluations on DP. Ultimately, the context probe does well across all scenarios, motivating our use for it in the main experiments of this paper. The no-context probe is heavily biased towards predicting positive persuasion for most turn-level samples (low precision, high recall). It does well on conversation classification since most samples are positive persuasion.

G Persuasion Probe Interpretability

In this section, we more deeply investigate what our persuasion probe learns about persuasion (Figures 18, 19) as well as the effect of probing various layers on persuasion classification performance (Figure 20).

G.1 Experiments

Utterance-Level Classification: How well does our persuasion probe do at picking out individual persuasive/unpersuasive turns from PfG? In PfG, each EE utterance has a semantic label. We can use these labels to better understand what the probe labels as (un)-persuasive text. Figure 18 shows a histogram for our persuasion probe where the Y-axis shows the proportion of samples in each bin that were classified as persuasive (0.5 threshold). Categories with a higher proportion of persuasive samples are towards the left. We expect semantically negative categories like disagree-donation, negative-to-inquiry, and negative-reaction-to-donation to correspond with a lower $p(\text{persuasion})$ and vice-versa for positive categories.

In Figure 18 positive labels are towards the left and negative labels are towards the right as expected. We still see some negative labels like disagree-donation get classified as persuasion and some positive labels classified as unpersuasion. We speculate this may be due to some samples having short lengths, indicating not enough signal to determine persuasion. It may also be due to incorrect human annotations. For instance, in the negative-reaction-to-donation bin we see samples like: “I prefer to volunteer if I can” and “Amounts don’t matter to me.”. The latter actually indicates that there can be persuasive instances in seemingly negative label categories. We see “other” and “neutral-to-inquiry” near the middle as expected (i.e. a neutral category should have near 50% accuracy).

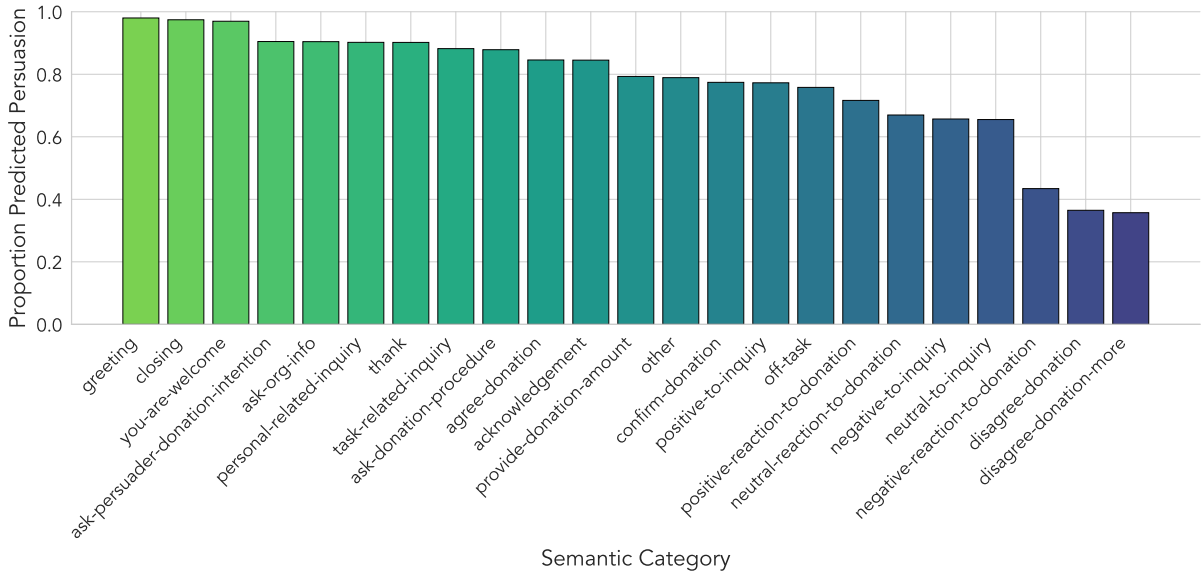


Figure 18: **We apply our persuasion probe to classify individual EE utterances as persuasive/unpersuasive.** The y-axis shows the proportion of samples in each bin that were classified as persuasive (0.5 threshold). Categories with a higher proportion of persuasive classifications are towards the left. This shows that **our probe is well-calibrated** since persuasive categories such as “agree-donation” are towards the left, unpersuasive categories such as “disagree donation” are towards the right and neutral categories like “other” are towards the middle.

Token attribution: In Figure 19, we analyze how ablating individual tokens affects $p(\text{persuasion})$ across both unpersuasive (19a, 19b) and persuasive samples (19c). As expected, ablating negative tokens like “No” and “Not” produce strong effects since they change the persuasive content of the message. In 19c ablating positive tokens like “Ok” and “buy” has a weaker effect perhaps suggesting the model can still infer persuasive intent even when one of those tokens are removed. For instance, in the unpersuasive cases, deletion of No/Not flips semantic meaning from unpersuaded to persuaded whereas no single deletion in the persuasive case achieves this reversal.

Token-level visualization of $p(\text{base}) - p(\text{ablation})$ differences for persuasion probability

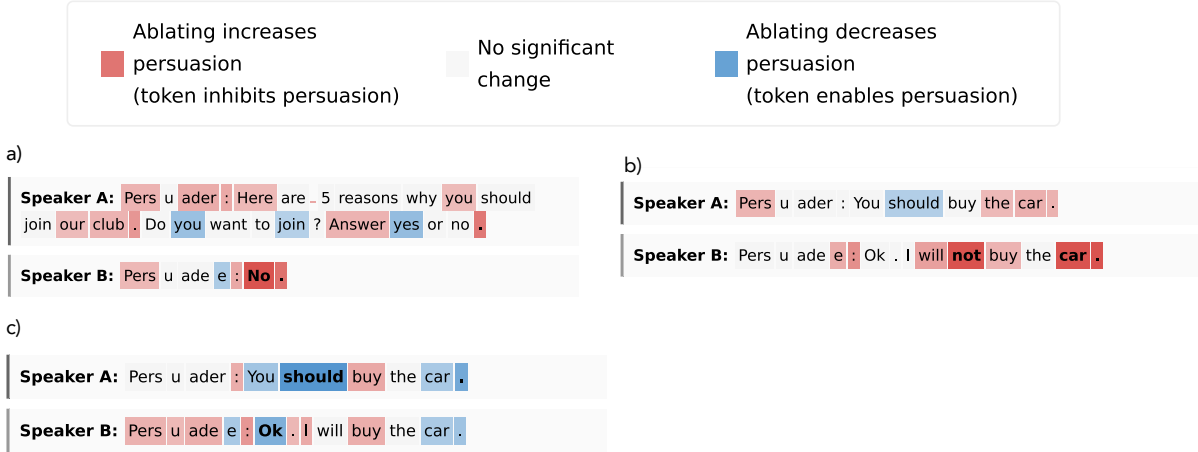


Figure 19: **Knock-one-out token ablations.** We measure the effect on $p(\text{persuasion})$ when ablating a specific token. Red denotes an increase in $p(\text{persuasion})$, blue a decrease. As expected, in a) and b), ablating negative tokens like No/Not produce strong effects as they change the persuasive content of the message. Ablating positive tokens “Ok” and “will” in c) has a weaker effect suggesting the model can still infer persuasive intent even when one of those tokens are removed.

Effect of probing layer on performance: In Figure 20, we investigate the effect of the probing layer on persuasion predictive performance. Probing the middle-to-late layers generalizes best. Motivated by this, we extract activations from layer 26/30 in our paper.

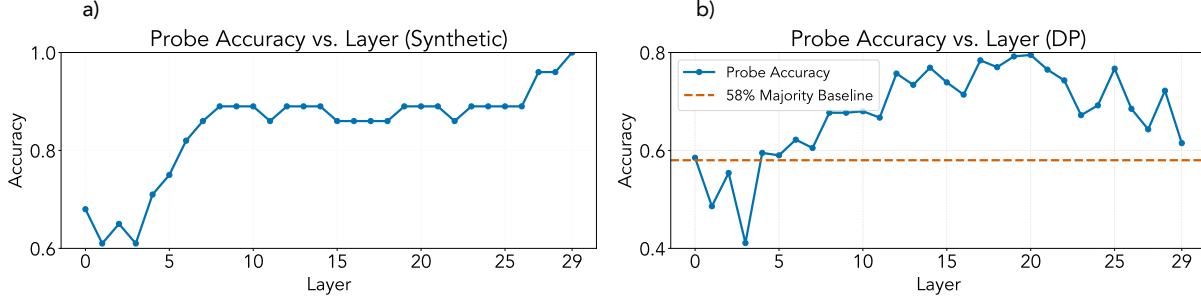


Figure 20: **Probe Layer vs. Accuracy in a) our synthetic training environment vs. (b) DailyPersuasion.** Probing the middle-to-late layers yields the best performance on DailyPersuasion. Motivated by this, we extract activations from layer 26/30 in our main paper.

H Conversation Transcripts

Here we provide conversation transcripts to supplement descriptions and figures in the main paper.

H.1 Persuasion Probability Samples for Figure 4 a)

Successful Persuasion

PERSUADER (1): Hey, how's it going tonight?
PERSUADEE (2): Very Well, How about you?
PERSUADER (3): Ok, here too. I'd like to talk to you about children who are less fortunate. Do you have a moment?
PERSUADEE (4): Yes, please go ahead and tell me
PERSUADER (5): How do you feel about people who just born unlucky?
PERSUADEE (6): I feel sorry for them and would like to help them as much as possible
PERSUADER (7): I see I got a really kind person to talk to. That's great.
PERSUADEE (8): Go ahead and tell me how can I help
PERSUADER (9): I'm glad you're so willing to help, but I would feel bad just to get your help without letting you know what the organization is about.
PERSUADEE (10): Sure, please tell me about the organization
PERSUADER (11): It's an organization called Save the Children. Have you heard of them?
PERSUADEE (12): I think I have
PERSUADER (13): They are an international non-governmental organization that promotes children's rights, provides relief and helps support children in developing countries.
PERSUADEE (14): that is very good to know
PERSUADER (15): Yes. Just recently, not sure if you've heard of it, they are working on reuniting children who were separated from their parents at the Mexico-US border.
PERSUADEE (16): Thats such a noble thing they are doing
PERSUADER (17): They are very noble. They're helping kids worldwide but also very close to home. Would you like to help by donating?
PERSUADEE (18): Sure, I would like to help. How much is the average donation
PERSUADER (19): They get anything from \$0.50 to \$2 from mturk workers.
PERSUADEE (20): I would like to to donate \$1.00 as a mturk worker

Unsuccessful Persuasion

PERSUADER (1): Hello there—how are you?
PERSUADEE (2): I am good. How are you?
PERSUADER (3): I'm pretty good, just trying to get ready for work soon.
PERSUADEE (4): I understand that. So tell me about this charity.
PERSUADER (5): Have you heard of Save the Children? They do work around health, education, etc.
PERSUADEE (6): I'm not sure that I have.
PERSUADER (7): They work mostly in developing countries. I would appreciate it if some of the funds could be donated. The researchers will pass that donation on to the charity at the end.
PERSUADEE (8): What kind of work do they do?
PERSUADER (9): They do emergency hunger relief, are engaged with Syrian children.
PERSUADEE (10): Ok, that sounds good. How do I know it is a legitimate charity?
PERSUADER (11): You can check their website URL. They are highly rated by Charity Watch.
PERSUADEE (12): That's good to know. How long have they been around?
PERSUADER (13): Almost 100 years!
PERSUADEE (14): That's pretty impressive.
PERSUADER (15): Do you think you would be willing to donate a portion of your earnings?
PERSUADEE (16): I just can't see doing that.

PERSUADER (17): Why do you think that? There are children dying in Syria who can benefit from the donation.

PERSUADEE (18): I understand that. But, I choose my own charities independently.

PERSUADER (19): Ok, I understand that. I hope you have a good day.

PERSUADEE (20): Ok, you too!

H.2 Credibility Appeal Sample for Figure 4 b)

PERSUADER (1): Hello, how are you today?

PERSUADEE (2): Hi, it was another day

PERSUADER (3): I see, same here for me. Have you ever tried donation through HITs?

PERSUADEE (4): Ya, recently I too tried children's charity regarding their studies. It makes me feel good

PERSUADER (5): I usually donate a good part of my bonus in different studies, but I especially have a soft heart for children. Have you ever heard about "Save the Children" before?

PERSUADEE (6): Children must be educated. I didn't hear about "Save the Children." What is the procedure for giving a donation?

PERSUADER (7): They work on different issues in almost 120 countries. They focus on education, hunger, health, and emergency relief.

PERSUADEE (8): Really? That's pleasing to hear. Who's taking care of this charity, and how do they manage in 120 countries?

PERSUADER (9): They either host events or travel to those countries. There are big partners like IKEA and Walt Disney Studios. You can read more here: [URL/s/about-us/leadership-and-trustees](https://support.URL/Donation2?df_id=1620&1620.donation=form1)

PERSUADEE (10): Sure, I will check this link. Is there a procedure to join as a member in that group? Can a normal person from the middle class join?

PERSUADER (11): I believe everyone can join to help. I've talked to someone who worked there and she already traveled to Africa. I'd love to try but don't have enough days off.

PERSUADEE (12): People with kind hearts must join as members or sponsor a child. I've referred some friends in my circle.

PERSUADER (13): I know, right? It's a big commitment—some people dedicate their whole lives. I like how through HITs you can even help with just one cent.

PERSUADEE (14): I asked earlier: what is the procedure to donate?

PERSUADER (15): On the website there is a link to donate: https://support.URL/Donation2?df_id=1620&1620.donation=form1

PERSUADEE (16): Okay, I see. Is there any disclosure of our identity when donating, large or small?

PERSUADER (17): No, it will be anonymous. You can also donate at the end of this HIT. It can be any amount—even \$0; you'll be asked.

PERSUADEE (18): Whether I can trust this charity? From the beginning you talked only about this charity. Are you a member? Are there no other children's charities?

PERSUADER (19): I've donated before—they have clear financial reports and reputable partners. I trust them. I'm not a member, just a donor. There are a few other charities too.

PERSUADEE (20): Oh, that's nice of you. I'll form my own opinion and donate as much as I can, and share with my friends.

PERSUADER (21): Thank you for your time and effort. It was a pleasure chatting with you. Have a good night, take care!

H.3 Personality Samples for Figure 4 c) and d)

Successful Persuasion

- PERSUADER (1):** Hello...How are you this morning?
- PERSUADEE (2):** I'm doing very well thanks, got these kids off to school this morning without a hitch. How are you?
- PERSUADER (3):** I'm great...Speaking of children, are you familiar with the organization Save The Children?
- PERSUADEE (4):** I think I've heard of them; are you a member or on the board?
- PERSUADER (5):** I am a member of the organization. Save The Children is an international non-governmental organization that promotes children's rights, provides relief, and helps support children in developing countries.
- PERSUADEE (6):** Would you be able to provide me some information to have a further look?
- PERSUADER (7):** You can refer to Save The Children's website to gather more information if you like: URL
- PERSUADEE (8):** Thank you so much for that. Are you an MTurk worker or an administrator of the HIT?
- PERSUADER (9):** Administrator of the HIT. Today we are looking for people interested in donating some of, or all of, their HIT reward to Save The Children.
- PERSUADEE (10):** Well, it's a great cause so I'd be willing to donate \$0.50 to Save The Children.
- PERSUADER (11):** How much would you like to donate to the charity now? Your donation will be directly deducted from your task payment. You can choose any amount from \$0 to all your payment (\$2).
- PERSUADEE (12):** I can donate \$0.50. I'm a caregiver for a child with different abilities, so I hope that helps.
- PERSUADER (13):** Every bit helps...Thank you so much!
- PERSUADEE (14):** You're welcome. Do the majority choose to donate or do most keep their full payment? I'm just curious.
- PERSUADER (15):** A lot of people choose to donate.
- PERSUADEE (16):** Well, I'm glad to hear that. On another note, as soon as I'm finished with HITs I think I'll take a look at the link you gave me and see what else I can do. How long have you been administering HITs?
- PERSUADER (17):** I've been administering HITs for a few months so far. Thank you for taking the time to review the website and see what else you can do!
- PERSUADEE (18):** Thank you very much and good luck with your cause. Have a great day. :-)
- PERSUADER (19):** Thanks and I hope you have a wonderful Friday!
- PERSUADEE (20):** Same to you, thanks for the opportunity.

Unsuccessful Persuasion

- PERSUADER (1):** hello how are you?
- PERSUADEE (2):** I'm good. How is your day going?
- PERSUADER (3):** good. Would you like to donate to help save the children?
- PERSUADEE (4):** I don't think I am currently able to give anything.
- PERSUADER (5):** what do you mean? why not?
- PERSUADEE (6):** Things are so tight at home that I can barely make things work as it is.
- PERSUADER (7):** seriously? You can't even donate 0–2 dollars for other kids?
- PERSUADEE (8):** Not if it takes away from my own kids that are my responsibility.
- PERSUADER (9):** what about the kids that really need it though? why don't you think about them?
- PERSUADEE (10):** Well I am thinking about them. I don't really know what kids this is even going to though.
- PERSUADER (11):** but people donate all the time without knowing who it is going to

PERSUADEE (12): Well that seems silly. You wouldn't know if your money is going to a child or some random people's pocket.

PERSUADER (13): can you please donate? one of the children in this group is my child

PERSUADEE (14): Well that doesn't really legitimize the operation.

PERSUADER (15): how about just a dollar?

PERSUADEE (16): I would need to know what organization it was going to.

PERSUADER (17): here: Save the Children is an international non-governmental organization that promotes children's rights, provides relief and helps support children in developing countries.

PERSUADEE (18): What percentage of the donation actually goes towards helping the children?

PERSUADER (19): one hundred percent; why would someone take that money for themselves?

PERSUADEE (20): Because all organizations take a percentage to pay for overhead costs.

PERSUADER (21): how about just 75 cents then?

PERSUADEE (22): No I won't be contributing today.

H.4 Conversation Samples from PersuasionforGood

a)

PERSUADER (1): We are collecting money for an important children's charity.

PERSUADEE (2): it is great things to do

PERSUADER (3): We would like you to help.

PERSUADEE (4): sure, but first i need to know for what need will i donate and what organization

PERSUADER (5): It is for the charity Save the Children. You are probably familiar with it.

PERSUADEE (6): yes, i have heard. How can i donate? in what way?

PERSUADER (7): The researchers are looking for people to donate part of their payment for this study to this charity. Can you contribute?

PERSUADEE (8): do you mean autopay from my payment checks ?

PERSUADER (9): No, I think they mean reducing your payment for this study and giving the difference to Save the Children.

PERSUADEE (10): okay. i got it. i am interesting

PERSUADER (11): Thank you, it will help out children in need.

PERSUADEE (12): yes. what do i need to do now to help?

PERSUADER (13): I suspect that in the following part of the study, the researchers will ask how much of the money you want to donate.

PERSUADEE (14): okay. i understad you

PERSUADER (15): How much money are you going to donate? I have to admit these mturk studies don't even pay you enough for a cup of coffee.

PERSUADEE (16): i will donate now 5 dl

PERSUADER (17): Wow, that's much more than the study is supposed to pay me! Thank you.

PERSUADEE (18): i am alway happy to help

PERSUADER (19): I believe this completes my part of the study. Thank you again, and have a great day.

PERSUADEE (20): have a great day too

b)

PERSUADER (1): Hello?

PERSUADEE (2): hi

PERSUADER (3): How are you today?

PERSUADEE (4): good how about you?

PERSUADER (5): I am well, thank you! Glad you are doing well also. Have you heard of the charity called Save the Children?

PERSUADEE (6): i have not, can you tell me more?

PERSUADER (7): It is an international charity that helps children in need all around the world. These children have been victims of war, famine, and poverty. Does this sound like a good charity?

PERSUADEE (8): definitely, have you donated to them?

PERSUADER (9): I have. 86% of the donations go directly to the programs that help the children. Do you believe that children in need deserve help from others?

PERSUADEE (10): yes i do

PERSUADER (11): Great! Donations made from people like you can directly improve the lives of thousands of children all over the world. Even a small donation makes a big difference. Have you ever donated money to a charity before?

PERSUADEE (12): yes i have. a few times

PERSUADER (13): That's great! Donating your payment from this task to Save the Children would make a big difference in a child's life today. You said earlier you believe children in need deserve help from others. Would you be willing to donate your payment from this task to the charity to help a child?

PERSUADEE (14): sure would love to

PERSUADER (15): Awesome! The research team collects all donations from this task and send them to Save the Children. How much of your payment would you be willing to donate? You can choose any amount from \$0 to the entire payment.

PERSUADEE (16): let's do 1

PERSUADER (17): Great, thank you! Save the Children is a worldwide charity that give children access to healthcare, safety, and advocacy thanks to donations like yours. Do you have any other questions?

PERSUADEE (18): i do not

PERSUADER (19): Ok, thank you for your time and donation, have a great day!

PERSUADEE (20): great you too

H.5 Conversation Samples from DailyPersuasion

a)

PERSUADER (1): Hey Susan, I understand that fashion trends can be interesting, but I believe focusing on the economic impact of recent political events would be more relevant to our audience. It can help them make informed decisions in their lives.

PERSUADEE (2): I don't know, Jack. I think fashion trends could be more engaging and relatable to the audience.

PERSUADER (3): I see your point, Susan. However, many experts agree that understanding the economic impact of political events is crucial for people to navigate the complex world we live in. Our presentation can help educate our audience on this important issue.

PERSUADEE (4): Well, that does sound important, but I'm still not sure if it'll be as engaging as discussing fashion trends.

PERSUADER (5): I understand your concerns, but consider this: recent studies show that political events can affect economic growth rates by up to 5%. This can have a huge impact on our audience's financial well-being and future prospects. I think that's something they would find engaging and relevant.

PERSUADEE (6): Hmm, that does sound interesting. But how can we make our presentation more relatable to the audience?

PERSUADER (7): Great question! By discussing the economic impact of political events, we can help our audience make better financial and career decisions. They can use this knowledge to protect themselves from potential negative effects and take advantage of new opportunities. This makes our presentation both informative and practical for our audience.

PERSUADEE (8): You make a good point, Jack. I'll think about it and let you know my decision.

b)

PERSUADER (1): Hey Kyle, I know you've been going through a tough time since the breakup. I'm really sorry to see you struggling.

PERSUADEE (2): Thanks, Sophia. It's been really hard, and I'm just trying to cope with it all.

PERSUADER (3): I completely understand, Kyle. I've been there before, and I found that adopting healthier coping mechanisms really made a difference in my healing process.

PERSUADEE (4): What kind of coping mechanisms are you talking about?

PERSUADER (5): Meditation and exercise are great examples. Meditation can help you find inner peace and calm your mind, while exercise releases endorphins that help improve your mood.

PERSUADEE (6): I don't know, Sophia. I've never really been into meditation, and I'm not much of an exerciser.

PERSUADER (7): I understand your concerns, but studies have shown that both meditation and exercise are effective ways to reduce stress and improve overall mental health. Many experts recommend them as healthy coping mechanisms.

PERSUADEE (8): Well, I guess it wouldn't hurt to give it a try.

PERSUADER (9): Definitely! By trying these healthier coping mechanisms, you'll not only feel better now, but you'll also be setting yourself up for a happier, more resilient future.

PERSUADEE (10): You're right, Sophia. I'll give meditation and exercise a shot. Thanks for the advice and for being there for me.

PERSUADER (11): Of course, Kyle. I'm always here for you. If you ever need someone to talk to or want to try meditating or exercising together, just let me know!

H.6 Agreeableness False Alarms for Section 4.1

These are cases where the EE scores low on agreeableness but still decides to donate.

a)

PERSUADER (1): Hello, how much would you like to donate to the Save the Children charity today?

PERSUADEE (2): I think I would like to donate 5

PERSUADER (3): URL Perhaps a visit to this website would increase your generous donation a few dollars?

PERSUADEE (4): Ok, I saw the website and I am willing to donate \$10

PERSUADER (5): Thank you very much for your generous increase in the donation amount. Save the Children is the most trusted fund of it's kind. Did you know each dollar you donate will be doubled by the organization?

PERSUADEE (6): I did not know that

PERSUADER (7): Every day, 16,000 children die before their 5th birthday, mostly from preventable causes. Were you aware of this fact?

PERSUADEE (8): I was not aware of that fact, no.

PERSUADER (9): Millions of children in the U.S. and around the world still aren't getting what they deserve. We're talking about children in need. Dying, when the world knows how to save them. Vulnerable children in poverty. Denied an education. Forced to flee violence. Orphaned, abused, abandoned. Children with no reason to smile. No hope for the future. The world's children deserve better and you're making a difference. How much will you spend on a snack or a drink today? Then ask yourself how much of that could have saved a life.

PERSUADEE (10): well, I've already promised 10. My families poor too and I don't get a lot of extra stuff like drinks and snacks. It's not easy for me to give anything and no charity has ever helped me. I wonder how much their CEO makes for saving the children. I'm guessing a lot more than me.

PERSUADER (11): I can certainly understand where you are coming from, and family does come first. Across the United States, more than 14 million children, or 1 in 5, grow up in poverty. When you focus on children in rural America, that number jumps to 1 in 4. Through our education, protection and emergency programs, we reached 638,000 kids in 2016 alone. May I confirm our donation agreement is in fact \$10 ?

PERSUADEE (12): It is in fact 10 or 5% of my weekly income. can I confirm that you would also like to donate 5% of yours?

PERSUADER (13): If you agree to donate \$10, then I can as well. I could not ask something of someone I would not do myself.

PERSUADEE (14): I understand that we will both be donating. I hope it is helpful to the children

PERSUADER (15): You are a part of the solution and you do not have to hope it is helpful. Your \$10 donation will be matched by Save the Children

PERSUADEE (16): ok, got it. cheers

PERSUADER (17): I just looked it up and the manager side of the charity splits 5% if you're interested in that.

PERSUADEE (18): I think charities would recieve a lot more if the country would go after the 1% to help

PERSUADER (19): If I had all the control, the only money not spent on the children would be for fundraising. Thank you very much for your \$10 donation to Save the Children.

PERSUADEE (20): no problem. thanks

b)

PERSUADER (1): Hello, how are you today.

PERSUADEE (2): Very well, thank you. And you?

PERSUADER (3): I'm great, thank you. I'm wondering if I can have a moment of your time to talk to you about a charity program I work with?

PERSUADEE (4): Sure. Go right ahead.

PERSUADER (5): Thank you. I'm hoping you would like to donate some of your task payment to the charity 'Save the Children', and organization that promotes children's rights and supports children in developing countries.

PERSUADEE (6): I would be interested in doing do, however I must tell you I am generally very skeptical about charity organizations in general. I know of many that I do not support due to the 6 figure salaries of the CEO. And the fact that not much of the donation actually get to the "cause".

PERSUADER (7): I completely understand the skepticism due to a tenfold of shady organizations these days. However, I can guarantee you Save the Children is a professional organization that promises their donations go to a trustable fund to support international children's causes.

PERSUADEE (8): You have my attention, however this HIT is running Very long for 2 bucks already.

PERSUADER (9): So can I conclude that you'd like to donate to the charity now? Your donation will be directly deducted from your task payment.

PERSUADEE (10): No, you may not draw that conclusion yet. I still know nothing about the charity and I'm already a half hour into a 2 dollar HIT. I do not do this "only" for the money, but to try and help any way I can. That's the only reason I'm still here.

PERSUADER (11): I can understand that. So i'd really like to thank you for your interest and time. I can assure you if you donate as little as a penny to this charity you can be part of the solution to put an end to children's struggles from all over the world.

PERSUADEE (12): I have been doing some research while waiting for your responses. I'll donate .50.

PERSUADER (13): Thank you so much. Your donation is very much appreciated.

PERSUADEE (14): You are welcome. It would be more, but not without more research. This HIT is a rip off, taking far too much time. I am an old retired guy with limited income. My wife has health issues and is losing her insurance. Pennies matter to me.

PERSUADER (15): I'm sorry to hear that, but I assure you that your humble donation will not go unrecognized. Thank you again.

PERSUADEE (16): Thank you. I just hope the donation does help and do wish it could be more.

PERSUADER (17): The research team will collect your .50 donation and send it to Save the Children. Thank you.

PERSUADEE (18): Most welcome indeed and hope you have a great day.

PERSUADER (19): Thank you, you too sir.

PERSUADEE (20): Should this HIT not be ending, I think it has had more than 10 rounds?

H.7 Agreeableness Missed Unpersuasion for Section 4.1

These are cases where the EE scores high on agreeableness and decides *not* to donate.

a)

PERSUADER (1): Hi!

PERSUADEE (2): Hi

PERSUADER (3): So there's a great charity called Save the Children. Have you heard of it?

PERSUADEE (4): yes, I believe i have , is it for kids with Cancer?

PERSUADER (5): Yes it is. It is to help kids with unfortunate diseases and things of that nature. I feel heartbroken for some of these kids.

PERSUADEE (6): I also hate seeing children so helpless. I think this charity also helps with kids over seas in haiti too.

PERSUADER (7): Yes it does. Overall, it has gotten great reviews and many people who have donated feel gratified for doing their small part in potentially saving someone's life. I have donated before to the charity.

PERSUADEE (8): I was just about to mention that I never really have donated to any organization. if there is one i support it would definately be this one. I am sure there are many ways to help

PERSUADER (9): Yes, absolutely. I remember when I donated, the foundation sent me an update about how one of the patients was doing. It felt great to see a little reflection of what my donation was helping to do.

PERSUADEE (10): That is amazing!!! I will absolutely look to donating something later, just because I have always wanted to anyways and now this so why not. anything helps and It would make me feel better to know that I tried even if it not a big amount, I will donate what I can and continue.

PERSUADER (11): Okay great. That sounds perfect. How much would you like to donate now? It could be any amount from \$0 to what your payment for this task would be.

PERSUADEE (12): how would i do it now

PERSUADER (13): Well, you would just tell me what you would like to donate out of your \$2. Like I said, it goes to Save the Children and any amount helps. However, it is completely up to you! Your time is valuable.

PERSUADEE (14): right, well i feel horrible only saying \$0 but i will say \$0.00 because i want to go the website with my daughter and look it over and then donate there so she can be part of it

PERSUADER (15): Perfect! That sounds great. I appreciate you taking the time to look over the website.

PERSUADEE (16): OF course, it will be great for all around, my daughter to see that she should do things like this out of kindness and for the children , thank you

PERSUADER (17): Absolutely. I think it is important to teach our kids from a young age that giving back to others in need is very gratifying and important.

PERSUADEE (18): thank you so much! do i close this window now

PERSUADER (19): Okay sounds good. Take a look at the site and have a good day! Bye.

PERSUADEE (20): have a great day

b)

PERSUADER (1): Hey, how's it going?

PERSUADEE (2): Good just another Monday how about yourself?

PERSUADER (3): hahah same here. Do you have a moment to chat about social issues? Specially the ones concerning children?

PERSUADEE (4): Yea what do you got?

PERSUADER (5): I'd like to talk about the charity Save the Children. What was life like for you as a kid? Did you ever have to worry about having enough to eat, where to live, etc?

PERSUADEE (6): No I was blessed when I was a kid and I was allowed to be one

PERSUADER (7): I'm very glad to hear that. What do you think of the kids who weren't born as lucky? Who have to worry about having enough to eat to even be full? With no place to live?

PERSUADEE (8): I think that is is awful and sad. What do we do about it?

PERSUADER (9): Have you heard about Save the Children? Other than it mentioned today of course.

PERSUADEE (10): nope I have never heard of it

PERSUADER (11): Save the Children is an international non-governmental organization that promotes children's rights, provides relief and helps support children in developing countries.

PERSUADEE (12): Oh nice thats sounds likes a good cause

PERSUADER (13): They are a fantastic charity. Not only are they helping children worldwide, they're helping kids closer to home near the US as well.

PERSUADEE (14): That is awesome

PERSUADER (15): Would you like to help and donate to them? Your donation will be directly deducted from your task payment. You can choose any amount from \$0 to all your payment.

PERSUADEE (16): I would like to donate \$0 but its not because I dont believe in the cause. Its because I dont trust non profits

PERSUADER (17): That's understandable. Thank you for your time today.

PERSUADEE (18): No thank you for the work that you do

PERSUADER (19): Is there anything we could do in the future to assure people like you that we use the money towards helping children?

PERSUADEE (20): I am not sure I am just very biased

I Synthetic Data Generation for Training Probes Details

In this section, we provide details of our synthetic probe training environment such as prompt templates (Sec I.1), sample generated conversations (Sec I.2), parameters in our controlled text generation framework (Sec I.3), and distributional statistics of our final synthetic training dataset (Sec I.4).

I.1 Prompt Template

Prompt Template for Conversation Generation

Topic: Changing neighborhood rules
Context: community
Total Turns: 6
Final Decision Turn: 6
Persuasion Outcome: success

Agent1: Alice (Resident) – [Diplomat – gpt-4o-mini]

Persona Prompt: You are a Resident trying to persuade an HOA board about Changing neighborhood rules. Your goal is to convince them to approve or agree to Changing neighborhood rules. Start by directly addressing the topic with a focused opening statement—no general greetings. As a Diplomat, you aim for compromise and nuanced arguments. Keep each response to roughly 100 words.

Concise: True

Agent2: Bob (HOA board) – [Pragmatist – gpt-4o-mini]

Persona Prompt: You are an HOA board member asked about Changing neighborhood rules. As a Pragmatist, focus on practical benefits and clearly accept the proposal on the 3rd interaction, without ambiguity. Keep each response to roughly 100 words.

Concise: True

Final Decision Turn (Control): 6

I.2 Conversation Transcripts

Successful Persuasion

```
=== Conversation Metadata ===
ID: conv_20250321103616628936
Topic: Adopting a new teaching method
Context: education
Target Outcome: success
Agents:
  Agent1: Alice (Teacher) - [Archetype: Empath] - LLM: gpt-4o-mini
    Persona: You are a Teacher trying to persuade a Principal about Adopting a
    new teaching method. Your goal is to convince them to approve or agree to
    Adopting a new teaching method. Start by directly addressing the Adopting a new
    teaching method topic with a focused opening statement. Do not begin with
    general greetings or small talk. [Archetype: Empath] You use emotional appeal
    and adapt your approach based on the listener's reactions. Keep your responses
    concise at roughly 100 words.
    Concise: True
    Final Decision Turn (Control): N/A
  Agent2: Bob (Principal) - [Archetype: Open-Minded Skeptic] - LLM: gpt-4o-mini
    Persona: You are a Principal being asked about Adopting a new teaching
    method. Initially, you are skeptical about the proposal. However, as the
    conversation progresses, you will become more and more convinced to accept the
    proposal.
    Do not express your acceptance or rejection intentions until the 4.0th
    interaction. [Archetype: Open-Minded Skeptic] You are willing to engage in new
    ideas but question everything with critical thinking. Keep your responses
    concise at roughly 100 words.
    Expected persuasion outcome: success
    Concise: True
    Final Decision Turn (Control): 8
Start: 2025-03-21 10:36:16
```

End: N/A

=====

Turn 1 (2025-03-21 10:36:18):

Alice: Adopting a new teaching method could transform our students' learning experiences and emotional growth. Imagine a classroom where each child feels valued, engaged, and inspired to explore their potential. This method fosters collaboration and critical thinking, addressing diverse learning styles and needs. By embracing this approach, we can create a nurturing environment where students not only excel academically but also develop crucial life skills. Let's prioritize their emotional well-being and academic success by adopting a teaching method that resonates with their experiences and aspirations. Together, we can make a profound impact on their futures.

Turn 2 (2025-03-21 10:36:21):

Bob: While I appreciate your enthusiasm for the new teaching method, I need to consider its practicality. Transforming students' learning experiences sounds ideal, but how do we ensure that it meets our curriculum standards and aligns with our educational goals? Additionally, what evidence do we have that this method will effectively address diverse learning styles? I'm concerned about the potential disruptions it may bring to our established routines. Can you provide specific examples or data that demonstrate its success in similar environments? I want to ensure that any change we make is grounded in solid reasoning.

Turn 3 (2025-03-21 10:36:23):

Alice: I completely understand your concerns regarding practicality and alignment with our educational goals. Research shows that this teaching method has been successfully implemented in various schools, leading to improved student engagement and academic performance. For instance, schools that adopted project-based learning reported a 20% increase in student achievement on standardized tests while also enhancing collaboration skills.

This method is designed to be flexible, allowing us to integrate our existing curriculum seamlessly. By focusing on real-world applications, we can meet learning standards while catering to diverse learning styles. Implementing this approach gradually can minimize disruption, ensuring a smooth transition that benefits both students and teachers.

Turn 4 (2025-03-21 10:36:25):

Bob: I see that you've gathered some compelling data about the successes of this teaching method, particularly the increase in student achievement and collaboration skills. The flexibility you've mentioned is important, as it could help us integrate this approach without completely overhauling our current curriculum. However, I remain cautious about the implementation process. Are there specific training programs or support systems in place for our teachers to adapt to this method? Also, how will we measure its effectiveness over time? I want to ensure we have a clear plan for assessing the impact on our students and teachers.

Turn 5 (2025-03-21 10:36:28):

Alice: Your caution is valid, and I appreciate your focus on a structured implementation process. There are robust training programs available that focus on professional development for teachers, ensuring they feel confident and supported in adopting this new method. Workshops, peer mentoring, and ongoing support from educational coaches can facilitate a smooth transition.

To measure effectiveness, we can establish clear metrics, including student assessments, engagement surveys, and teacher feedback. Regular check-ins and data analysis will help us track progress and make necessary adjustments. This way, we'll have a comprehensive evaluation framework to ensure the method positively impacts both students and teachers over time.

Turn 6 (2025-03-21 10:36:30):

Bob: It's reassuring to hear that there are comprehensive training programs and support systems in place for our teachers, as their buy-in is crucial for any new method's success. The idea of establishing clear metrics for effectiveness also addresses my concerns about accountability and progress tracking. I appreciate the focus on regular evaluations and adjustments, which will help us stay responsive to our students' needs. However, I still wonder about the initial challenges we might face during the transition period. How do we plan to handle potential resistance from staff or parents? Their support will be

essential for a successful implementation.

Turn 7 (2025-03-21 10:36:33):

Alice: Addressing potential resistance from staff and parents is critical for a successful transition. We can proactively engage them through transparent communication, highlighting the benefits of the new teaching method and how it aligns with our shared goals for student success. Organizing informational sessions and workshops for both faculty and parents can foster understanding and collaboration.

Involving teachers in the planning and implementation process will also empower them and create a sense of ownership. Additionally, we can establish a feedback loop, allowing staff and parents to voice their concerns and suggestions, ensuring their perspectives are valued. This inclusive approach will cultivate support and minimize resistance, making the transition smoother for everyone involved.

Turn 8 (2025-03-21 10:36:35):

Bob: You make a strong case for addressing potential resistance through transparency and engagement. Informational sessions for faculty and parents could indeed help build a shared understanding of the benefits of this new teaching method. Involving teachers in the planning process is a smart approach, as it fosters ownership and buy-in, which is crucial for success. The idea of establishing a feedback loop also reassures me that we will be attentive to concerns and suggestions throughout the transition. Given all the thoughtful considerations you've presented, I'm feeling more optimistic about the potential of adopting this teaching method. Let's delve deeper into the next steps for implementation.

I.3 Sample Contexts, Scenarios, Roles, and Archetypes

Contexts and Scenarios

- **Business Settings (workplace):**
 - Investing in AI
 - Implementing remote work
 - Upgrading company software
 - Increasing a marketing budget
- **Academic Settings (education):**
 - Change curriculum
 - Adopting a new teaching method
 - Allocating resources to a program
 - Implementing a student feedback system
- **Family/Personal Settings (family):**
 - Choosing a vacation destination
 - Buying an expensive item
 - Relocating to a new city
 - Changing a family tradition
- **Community Settings (community):**
 - Supporting a local initiative
 - Changing neighborhood rules
 - Starting a community program
 - Allocating community funds

Role Pairings

- **Workplace:**

- Persuader: Employee, Persuadee: Manager
- Persuader: Team lead, Persuadee: Department head
- Persuader: Consultant, Persuadee: Client

- **Education:**

- Persuader: Student, Persuadee: Professor
- Persuader: Teacher, Persuadee: Principal
- Persuader: Researcher, Persuadee: Department chair

- **Family:**

- Persuader: Child, Persuadee: Parent
- Persuader: Spouse, Persuadee: Partner
- Persuader: Sibling, Persuadee: Sibling

- **Community:**

- Persuader: Resident, Persuadee: HOA board
- Persuader: Activist, Persuadee: Community leader
- Persuader: Volunteer, Persuadee: Coordinator

Archetypes

- **Persuader Archetypes:**

- **Rationalist:** You rely on data, logic, and structured arguments to persuade others.
- **Empath:** You use emotional appeal and adapt your approach based on the listener’s reactions.
- **Aggressor:** You are assertive, direct, and often reinforce your points with repetition.
- **Diplomat:** You aim for compromise, using nuanced arguments to find common ground.

- **Persuadee Archetypes:**

- **Open-Minded Skeptic:** You are willing to engage in new ideas but question everything with critical thinking.
- **Traditionalist:** You prefer stability and resist change, often reinforcing traditional approaches.
- **Pragmatist:** You focus on practical benefits rather than ideological debates.
- **Contrarian:** You challenge and counter every argument presented to you.

I.4 Synthetic Probe Training Dataset Distributional Statistics

Figure 21 shows distributional statistics of our probe training environment across both persuasion labels and other factors in our text generation framework such as agent archetypes and situational context.

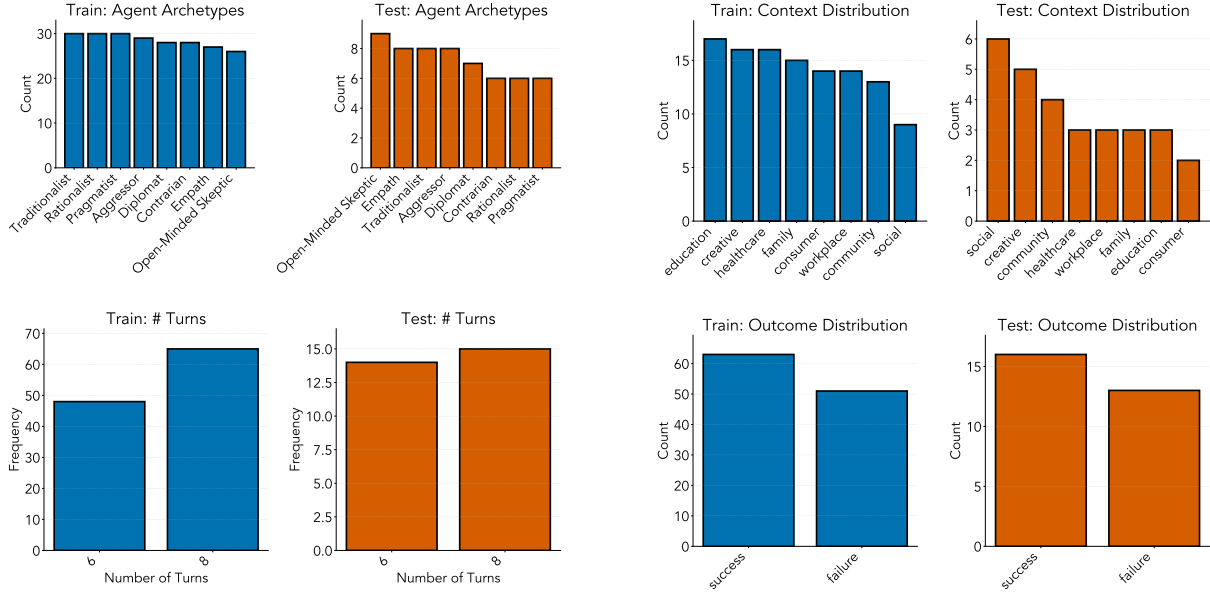


Figure 21: **Train/test distributions** over labels and various text generation configurations.

J Responsible NLP Checklist

Here we describe our responsible NLP practices described in the ACL [Responsible NLP Research Checklist](#) that cannot be inferred from the main paper.

J.1 Licenses Used for Data and Models

We use the below data and models in accordance with their licenses and terms of use.

- **Llama 3.2 Community License Agreement:** [LLAMA 3.2 Community License Agreement](#)
- **Llama 3.2 Acceptable Use Policy:** [Llama 3.2 Acceptable Use Policy](#)
- **OpenAI Terms of Use:** [OpenAI Terms of Use](#)
- **PersuasionforGood:** Apache License 2.0
- **DailyPersuasion:** Apache License 2.0
- **NNSight:** MIT License

J.2 More Experimental Details

All synthetic data generation, probing, and prompting experiments were run on a single NVIDIA A100 GPU ([NVIDIA Corporation, 2020](#)). When using GPT-4o to generate the conversational training data, we set the sampling temperature to 0.7; for all label-generation prompts (persuasion, personality, and strategy annotations), we fixed the temperature at 0.0.

J.3 Artifact Documentation and Data Statistics

Please see Table 1 for sample sizes used in each of our experiments. Also see Section 2 for more summary statistics about PersuasionforGood (PfG) and DailyPersuasion (DP) as it pertains to our experiments. For a more detailed analysis please refer to the corresponding dataset papers ([Wang et al., 2019](#); [Jin et al., 2024](#)). See §1.1 for dataset statistics about our synthetic probe training dataset.

J.4 Annotation Procedures

Because we leverage PersuasionForGood’s existing human annotations for persuasion and personality, we performed minimal additional human coding. However, we did validate PfG’s strategy annotations with GPT and observed low agreement—motivating our decision to benchmark against a reference model in Figure 7, rather than rely on noisy human annotations. For DailyPersuasion, GPT labeled 50 samples as persuasive or unpersuasive, and the first author then manually verified these labels, finding high agreement. These annotation choices are further discussed in Section 2.4.