HAL: Inducing Human-likeness in LLMs with Alignment

Anonymous ACL submission

Abstract

Aligning language models to qualitative behavioral traits, such as human-likeness, remains difficult because they are hard to define, measure, and optimize. As a result, improvements in human-like behavior are largely driven by scale or broad supervised training, rather than targeted alignment. We introduce Human Aligning LLMs (HAL), a framework for aligning language models to conversational human-likeness using an interpretable, data-driven reward. HAL derives explicit conversational traits from contrastive dialogue data, combines them into a compact scalar score, and uses this score as a transparent reward signal for alignment with standard preference optimization methods. Using this approach, we align models of varying sizes without affecting their overall performance. In large-scale Chatbot Arena-style human evaluations, a model aligned with HAL is more frequently perceived as human-like in conversation. Because HAL operates over explicit, interpretable traits, it enables inspection of alignment behavior and diagnosis of unintended effects. More broadly, HAL demonstrates how soft, qualitative properties of language–previously outside the scope for alignment–can be made measurable and aligned in an interpretable and explainable way.

1 Introduction

Human communication is the product of millions of years of social evolution, shaped by subtle and largely unspoken norms. While these norms are difficult to articulate; however, once broken, easily detected. When artificial agents fail to reproduce them, interactions can feel mechanical or uncanny.

Human-like conversational behavior is especially important in settings where interaction quality matters more than task completion. These include role-play and character simulation (Shanahan et al., 2023; Wang et al., 2024; Tao et al., 2024; Chen et al., 2024), communication training (Yang et al., 2024; Burgues et al., 2024; Hasan et al., 2023), patient simulation in healthcare and mental health contexts (Louie et al., 2024; Haut et al., 2025; Elhilali et al., 2025; Baseman et al., 2025; Scherr et al., 2023), and others.

Despite its importance, conversational human-likeness remains difficult to define and even harder to measure. Humans can often tell whether a conversational partner is human or artificial, but this judgment is typically holistic and implicit rather than based on explicit criteria. As a result, there has been no systematic way to measure human-likeness, and consequently no clear reward signal for aligning models toward it. This has left training language models to be more human-like largely out of reach for alignment research.

Recent large-scale Turing test results highlight both progress and limitations (Jones and Bergen, 2025). `GPT-4.5` was judged to be human in 73% of comparisons, while `LLaMA-3.1-405B` achieved near-chance performance and smaller baselines performed far worse. These findings suggest that scale can improve perceived human-likeness, but they do not explain why, nor do they provide a clear recipe for training models to be more human-like.

In this paper, we introduce *Human Aligning LLMs* (HAL), a framework for quantifying conversational human-likeness and using it as a reward for alignment. Our approach is entirely data-driven: we extract recurring human-likeness cues from contrastive dialogue data (e.g., Turing tests), compress them into a compact and interpretable set of traits, and combine them into a single scalar score. We then use this score as a reward signal for alignment with standard preference optimization methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2023). Across models of varying sizes, we show that alignment with HAL leads to clear improvements in perceived human-likeness under human evaluation, while largely preserving performance on other benchmarks.

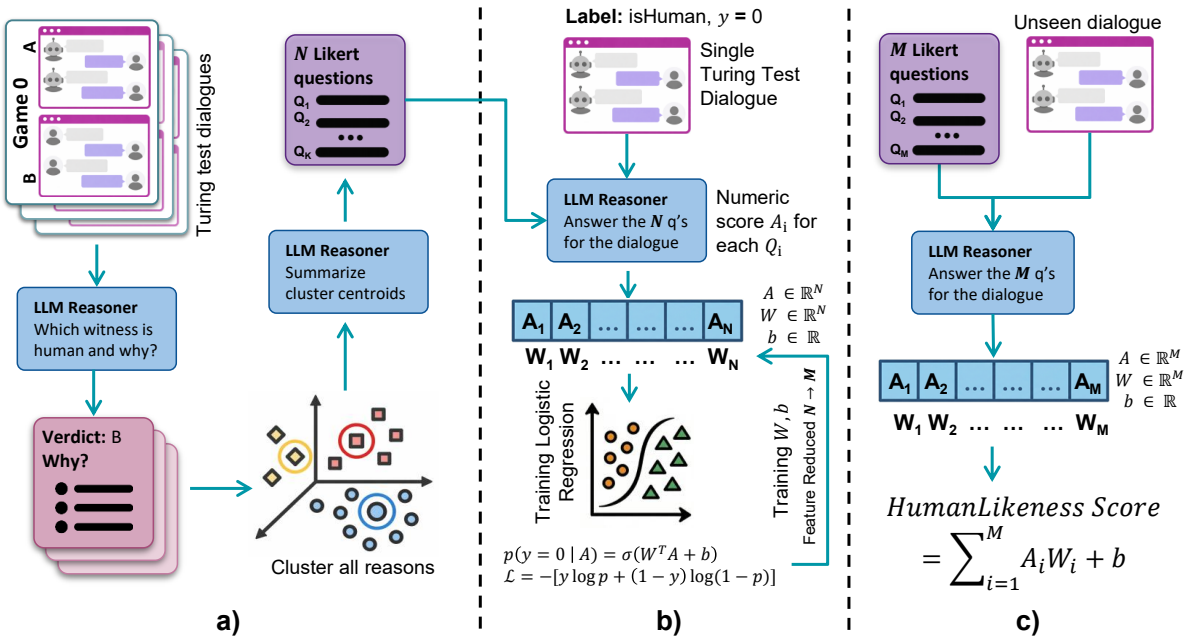

Figure 1: The HAL pipeline: (a) identifying human-likeness traits from contrastive dialogues (e.g. Turing tests), (b) learning trait weights via a proxy classification task, and (c) computing a human-likeness score for alignment.

Concretely, our proposed framework HAL:

1. identifies recurring conversational traits that reliably distinguish human–human from human–AI dialogue,

2. compresses these traits into an interpretable, scalar measure of conversational human-likeness,

3. uses this measure as a reward signal for alignment with standard preference optimization methods, and

4. demonstrates through human evaluation that models aligned with HAL are more frequently perceived as human-like.

More broadly, HAL offers a general methodology for inducing soft, qualitative traits in language models—traits that are difficult to specify directly, but can be inferred from contrastive data. By making such traits measurable and interpretable, this work opens new possibilities for controllable, transparent, and human-centered alignment beyond conventional objectives.

Our best weights are available on HuggingFace and Ollama. Our training code and datasets are available at: `https://anonymous.4open.science/r/hal-0B35`

2 What Makes Human Conversation Human?

We begin with a simple premise: to make a model more "human-like" in conversation, we should first understand which conversational cues best help identify it. Once we are able to understand and quantify that, we can align the model to demonstrate more of that behavior.

2.1 Characteristics of Human-likeness From Turing Tests

The Turing test is an imperfect proxy for human-likeness, but it provides two ingredients that are difficult to obtain otherwise: paired dialogues that are designed to be compared, and which side is human. We analyze the Turing test dialogue transcripts released by Jones and Bergen (2025).

The original dataset consists of 1116 games, where the *investigator* I interacts with two unknown conversational partners (the *witnesses* W's), and then decides which W is human. Each game consists of two conversations between I and two W's, along with (i) the ground truth, (ii) the investigator's decision, and (iii) a brief free-form explanation of the decision. We filter out games where either conversation is < 50 words, as short conversations are either obvious or uninformative. Hence, we are left with 557 games, each consisting of a human-AI and a human-human conversation.

In our filtered dataset, the human judge has an accuracy of 54.58% in identifying correctly who is human, which is slightly better than random.

LLMs as Turing judge. A natural starting point is to treat the investigators' free-form explanations as a source of human-likeness cues. In practice, they are inconsistent in format and very frequently incorrect. We therefore construct an *LLM-based Turing judge* that evaluates the same paired dialogues, but produces explicit and structured reasons. For each Turing test game, we pass the judge with two dialogues in random order, ask it to (i) predict which witness is human, and (ii) provide 3-5 Likert-style statements that helped make the decision on this specific pair of dialogues (full prompt at Appendix 9).

The goal of this experiment is not to claim a new best Turing judge, but to identify a set of high-signal descriptions of human-likeness cues using the classification accuracy as a proxy for the quality of the reasons. We evaluate a cohort of commercial and open-source models as displayed in Table 2. In our experiments, `GPT-5` with `high` reasoning attained the highest accuracy of 64.81%, and therefore its "reasons" are used for further analysis.

Creating a compact set of characteristics. Running the LLM-judge over the dataset yields $2,735$ total natural-language reason statements, many of which are redundant. We embed each reason statement using a sentence encoder (Wang et al., 2020) and cluster them using a density-based clustering algorithm (Campello et al., 2013). We find 53 representative clusters and extract their centroid. The 53 centroid statements still contained redundancies. Hence, we further instruct an LLM to summarize the 53 clusters into distinct Likert-style statements (prompt Appendix Figure 7). This yields a final inventory of 32 characteristics, presented in Appendix Table 5. Henceforth, we refer to these 32 traits as *Human-Like 32 Questions* or HL32Q.

The resulting characteristics reflect what repeatedly distinguishes human and model dialogues in this dataset under a Turing-style comparison. This pipeline is visualized in Figure 1 a).

3 Quantifying Human-likeness

In previous section, we identified a compact set of conversational traits (HL32Q) that repeatedly distinguish human–human from human–AI dialogue in a Turing-style setting. Our next goal is to turn these qualitative traits into a quantitative signal and derive a single score to a dialogue that reflects how human-like it appears. This simple score will be used as a reward signal in alignment training. To extract quantitative signals from qualitative statements, we draw on survey methods developed in social science and psychology.

3.1 Human-likeness Classifier

Given a dialogue, an LLM judge (*HL32Q Judge*) rates its agreement with each of the 32 statements only based on the witness responses on a 1–5 Likert scale (full prompt Appendix Figure 10). Unlike the Turing test setting, this is not a pairwise comparison. Each dialogue is scored independently, producing a fixed-length feature vector $\mathbf{A} \in \mathbb{R}^{32}$. This representation compresses a dialogue into a small number of high-level and complex conversational cues.

We propose a simple method to compress this rich vector $\mathbf{A}$ into a single scalar human-likeness score. We train a logistic regression model that takes $\mathbf{A}$ as input and predicts y, indicating whether the witness in the original dialogue is human ($y = 1$) or an AI ($y = 0$). By training with backpropagation, we learn the weight matrix $\mathbf{W}$ and add bias $\mathbf{b}$. The model learns a linear decision boundary over the 32 features, yielding a weight for each trait that reflects its contribution to distinguishing human from AI dialogue. Even though more accurate aggregation methods can be developed, we focus on simplicity and interpretability in the current study.

Formally,

$$\begin{gathered}\mathbf{A} \in \mathbb{R}^N, \quad \mathbf{W} \in \mathbb{R}^N, \quad b \in \mathbb{R}, \quad y \in \{0,1\} \\ p(y = 1 \mid \mathbf{A}) = \sigma(\mathbf{W}^\top \mathbf{A} + b) \\ \mathcal{L} = -\big[y \log p + (1-y)\log(1-p)\big]\end{gathered} \tag{1}$$

where $N = 32$ and $\sigma(\cdot)$ denotes the logistic function.

In 10-fold cross-validation repeated over 20 random splits, this simple linear model achieves 77.47% accuracy when using `GPT-5` as the HL32Q judge (Table 2).

3.2 Feature Reduction and Single Score of Human-likeness

Alongside high accuracy in distinguishing human dialogues from AI, we wish to make the features simple and interpretable. We therefore select the top $M = 16$ traits ranked by absolute weight magnitude $|W_i|$. Using only these features reduces ac-

No.	Statement	Weight
Q1	Keeps replies brief and casual without over-explaining.	1.3736
Q2	Uses emojis, emoticons, and playful elongations.	-0.2474
Q3	Makes niche cultural references from personal memory and assumes shared context.	-0.5006
Q4	Uses lowercase texting style.	0.4703
Q5	Shows small typos, uneven punctuation, and informal grammar typical of quick texting.	0.7079
Q6	Builds on the other person's message and context.	0.3124
Q7	Uses natural, idiomatic phrasing.	-0.7266
Q8	Shows reciprocity by asking natural, context-aware follow-up questions that advance the chat.	-0.4266
Q9	Uses casual, playful humor.	-0.3120
Q10	Admits not knowing and asks to learn instead of inventing details.	0.1217
Q11	References immediate context or recent activity.	-0.3562
Q12	Uses casual slang, abbreviations, and shorthand naturally.	-0.2189
Q13	Explains choices with simple personal reasons and constraints.	0.3429
Q14	Stays on topic and steers the conversation rather than mirroring or deflecting.	-0.1819
Q15	Sometimes shows impatience and ends the chat quickly with a brief nicety.	0.2563
Q16	Gives direct answers about self with concrete personal details.	-0.1905

Table 1: HL16Q: Selected 16 Likert-style statement and their weights W found by logistic regression. Bias $b = -2.662$.

curacy only marginally, to 77.12% while giving a notable boost in interpretability and explainability.

We refer to this reduced set as *Human-Like 16 Questions* (HL16Q). After retraining the logistic regression on the full Turing test dataset using these 16 features, we fix the learned weights $\mathbf{W}$ and bias b. Table 1 lists the selected statements and their learned weights. The signs and magnitudes reflect how each trait shifts the model toward or away from a human classification in this dataset; however, they are not a universal measure of human-likeness.

This method allows us to define a single scalar score for any new dialogue:

$$\mathbf{A} \in \mathbb{R}^M, \quad \mathbf{W} \in \mathbb{R}^M, \quad b \in \mathbb{R}$$
$$\text{HumanLikeness}(\mathbf{A}) = \sum_{i=1}^{M} A_i W_i + b \tag{2}$$

where $M = 16$.

We refer to this value as the *HL16Q score*. Higher scores indicate more human-like conversational behavior under this metric.

	Reasoning	Pairwise	Accuracy (%)
Live Turing test			
Human judge	-	Yes	54.58
Finding characteristics			
GPT-4.1	-	Yes	53.68
GPT-4.1-mini	-	Yes	46.14
GPT-5	high	Yes	**64.81**
GPT-5-mini	high	Yes	53.32
GPT-OSS:120B	high	Yes	40.41
GPT-OSS:20B	medium	Yes	38.73
HL32Q Judge			
GPT-4.1	-	No	70.51
GPT-4.1-mini	-	No	67.45
GPT-5	high	No	**77.47**
GPT-5-mini	high	No	73.77
GPT-OSS:120B	high	No	73.92
GPT-OSS:20B	high	No	70.56
HL16Q Judge			
GPT-5	high	No	77.12

Table 2: Finding characteristics aim to identify the differences between human–human and human–AI data and generate plausible reasons for these differences. The HL32Q judge aims to determine optimal weights for calculating a numerical human-likeness score. Accuracy on the filtered Turing test dataset from Jones and Bergen (2025) serves as a proxy for both tasks.

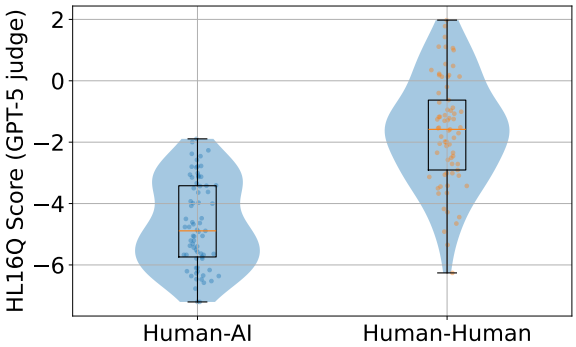

Figure 2: Violin plot of HL16Q Score on Out-of-distribution (OOD) dataset containing Human-AI and Human-Human conversations.

3.3 Evaluating on OOD Data

Finally, we test whether the HL16Q score generalizes beyond the Turing test data used to derive it. We evaluate on an out-of-distribution dataset consisting of 73 human–human and 73 human–AI dialogues from a separate cancer communication study (Haut et al., 2025). This dataset differs in topic, style, and collection procedure (Data proxy in Appendix Table 6).

Figure 2 shows the distribution of HL16Q scores for the two groups. Human–human conversations receive substantially higher scores than human–AI conversations, with the distribution shifted upward, a higher median, and limited overlap. The mean score increases from -4.68 for human–AI conver-

sations to -1.66 for human–human conversations, yielding a mean difference of $\Delta = 3.02$ with a 95% bootstrap confidence interval of $[2.50, 3.54]$. This difference is statistically significant under a one-sided Mann–Whitney U test ($U = 4835.00$, $p = 1.01 \times 10^{-17}$). A large effect size of AUC $= 0.907$ further supports that HL16Q captures transferable cues of human-likeness rather than merely fitting the original Turing-test data.

4 Inducing Human-likeness with Alignment

Having derived a single, interpretable score for conversational human-likeness (HL16Q Score), we now use it as a reward signal for alignment. Our goal is to nudge models toward behaviors that helped distinguish human–human dialogue from human–AI.

We frame this as a preference learning problem. For a given conversational prompt, we generate multiple candidate dialogues, score them with the HL16Q Judge, and construct ranked pairs where the more human-like dialogue is preferred. We use these pairs to align models with Direct Preference Optimization (DPO) (Rafailov et al., 2023).

4.1 Persona Synthesis

To create diverse yet controlled conversational settings, we synthesize personas that serve as prompts for dialogue generation. We begin with 500 seed personas from the SynthLabs PERSONA dataset (Castricato et al., 2025). These seeds are split into 450 training personas, 25 test personas, and 25 personas reserved for human evaluation. All augmentation is performed after this split to avoid data leakage.

Each seed persona is expanded into four related personas with some overlapping traits. Gender is preserved, while age is perturbed by up to 5%. Appendix Figure 6 shows some demographic distribution of our generated data. This results in 1,800 training personas, and 100 personas each for testing and human evaluation. To avoid overly polite or agreeable behavior, we randomly assign a negative personality trait (e.g., anxious, hostile, arrogant, etc.) to 5% of personas.

For each persona, we generate a detailed biography using `GPT-4.1`. For consistency and ease of evaluation, we limit our generated dialogues on medical communication domain. We further fabricate a medical condition and a reason for a clinical visit, which together define the conversational context. The full prompt structure is provided in Appendix Figure 8. Appendix Table 7 shows two personas generated from the same seed persona side by side.

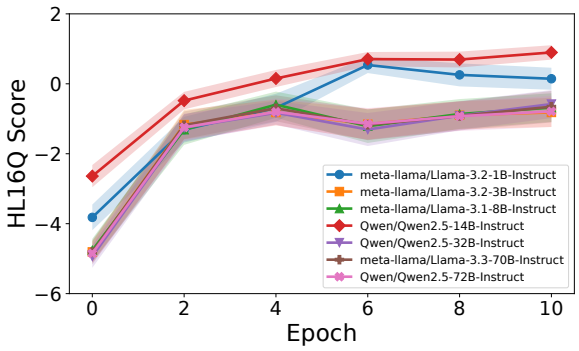

Figure 3: Human-likeness Score with 95% Confidence Interval (shaded) on 10 epochs of training with DPO for models of 7 incremental sizes. Test sampled every two epochs.

4.2 Dialogue Generation and Ranking

For each of the 1,800 training personas, we generate candidate dialogues using a diverse set of models: `GPT-4.1`, `GPT-4.1-mini`, `GPT-4.1-nano`, `GPT-5`, `GPT-5-mini`, `GPT-5-nano`, `LLaMA-3.1-405B`, and `Qwen2.5-14B`. We sample a model from this list 7 times and produce 7 dialogues per persona, resulting in 12,600 dialogues in total. Model statistics are summarized in Appendix Table 8.

Each dialogue is independently scored using the HL16Q judge (`GPT-5`). For each persona, the 7 generated dialogues yield 21 possible pairs. We retain only pairs whose HL16Q scores differ by at least $0.5\times$ the standard deviation across the dataset. Within each retained pair, the higher-scoring dialogue is labeled as *chosen* and the lower-scoring one as *rejected*. This filtering yields 7,175 ranked dialogue pairs, with all 1,800 personas represented.

4.3 Training

We fine-tune seven open-access models from multiple model families and generations, with parameter counts of 1B, 3B, 8B, 14B, 32B, 70B, and 72B, using DPO (Rafailov et al., 2023). The models were simple instruction-tuned models, with no Mixture-of-Experts (MoE) and Chain-of-Thought (CoT) training, as these add more complexity in alignment training. In a typical alignment pipeline, an alignment step is often followed by supervised fine-

tuning (SFT) to ensure the model adheres to a specific output format. We omit this intermediate supervised fine-tuning stage because the instruction-tuned base models already followed the required dialogue format under prompting. Furthermore, this helps us isolate the effect of HAL reward in preference optimization and test whether the pairwise preferences induced by HL16Q score are sufficient to shift conversational behavior to be more human-like.

We train all models for 10 epochs with DPO using $\beta = 0.1$, AdamW optimization in 8-bit precision, a learning rate of 5×10^{-5}, linear scheduling, and a 10% warmup. Training uses an effective batch size of 32 via gradient accumulation, a maximum sequence length of 1024, and no weight decay. We apply LoRA (Hu et al., 2022) with rank 16, $\alpha = 32$, and dropout 0.1. All models are trained with 4-bit quantization using HuggingFace Accelerate (huggingface.co, 2026) in data-parallel mode on 4 NVIDIA H100 GPUs. The total GPU time for the final run of all 7 models is 48.5 hours.

4.4 Training Validation

Across models of various model families and sizes, the alignment training consistently improves the HL16Q score over training epochs (Figure 3). With the exception of `LLaMA-3.2-1B`, most models maintain an upward trajectory in the 10th epoch, indicating there is more room for improvement.

Interestingly, we observe no clear relationship between parameter count and gains in human-likeness. Majority models follow similar training trajectories, regardless of size. `Qwen2.5-14B` and `LLaMA-3.2-1B` start from stronger initial scores and show larger absolute improvements, suggesting that initial conditions of model training may play a larger role than model scale in human-likeness training.

Despite these improvements, mean HL16Q scores remain negative for most models after training. We attribute this to a domain mismatch between the alignment data and the original Turing test data, for which the bias b was set.

4.5 Interpretation

Because the HL16Q score is a weighted sum of interpretable traits, we can inspect how alignment affects individual conversational characteristics. This allows us to diagnose behavioral changes during training and potential reward hacking.

Figure 4 shows the per-question score distributions for `Qwen2.5-14B` before and after alignment. The largest changes occur on Question 1, which has the highest weight and captures brief, casual responses. We also observe distributional collapse on several traits (e.g., Q1, Q8, Q14, Q16), where variance decreases after training. In contrast, other traits (e.g., Q11, Q12) exhibit increased spread.

While these observations are not conclusive, they illustrate the value of an interpretable reward. The HAL framework allows us to inspect how alignment reshapes specific conversational behaviors, rather than treating human-likeness as an opaque scalar objective.

5 Evaluating Human-likeness Training

To assess whether alignment with HAL leads to perceptible improvements in human-likeness, we conduct a controlled human evaluation. We focus on direct human judgments, using a Chatbot Arena–style A/B comparison, where participants interact with two randomly selected models side by side and decide which one feels more human in conversation.

5.1 Chatbot Arena

Evaluation setup. We use a Chatbot Arena–style pairwise comparison interface (Chiang et al., 2024). For each trial, participants interact with two chatbots shown side by side in random order (Appendix Figure 5). Both chatbots receive the same persona prompt, sampled from a held-out set of 100 personas. Each participant completes five trials with five unique, randomly selected personas. In each trial, the model pair is also randomly selected.

Participants must exchange at least two turns with each chatbot before rating which chatbot is more human-like using a five-point preference scale: *Certainly A*, *Likely A*, *Tie*, *Likely B*, or *Certainly B*. The user experience is identical across conditions except for the underlying model, allowing us to isolate the effect of model behavior. To reduce the risk of personally identifiable information (PII) disclosure and standardize the interaction, participants are instructed to respond as a unique fabricated clinician persona in each trial.

We evaluate three models: `Qwen2.5-14B (Base)`, `Qwen2.5-14B (HAL)`, and `GPT-4o-mini`. The Qwen models are hosted locally with Ollama on a server with two NVIDIA A6000 GPUs, while `GPT-4o-mini` is accessed through the OpenAI API.

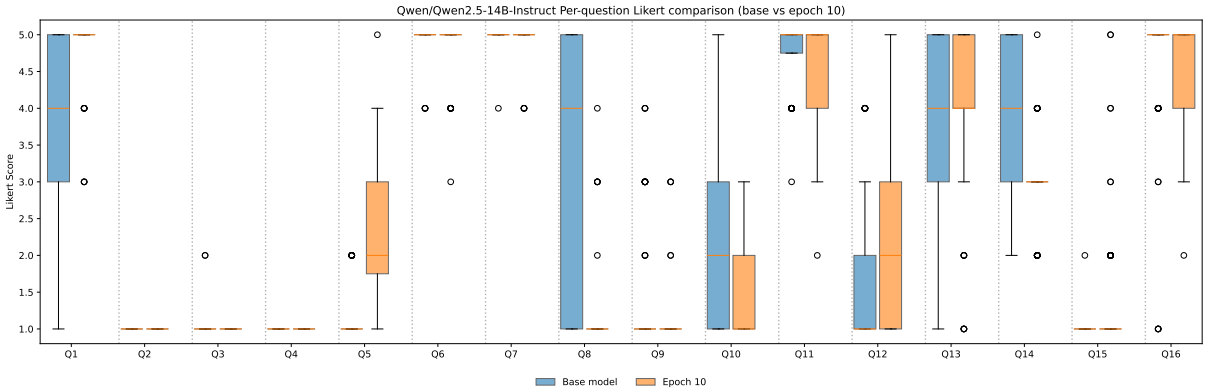

Figure 4: Qwen2.5-14B HL16Q individual statement interpretation

Participants. We recruit participants from Prolific[1] with the following criteria: located in the United States, fluent in English, at least a high school education, and a minimum of two prior Prolific submissions. Each participant can take part only once, and unusually fast submissions are automatically rejected. In order to proceed with the study, each participant was required to view and provide online agreement with the consent form. This study was reviewed by the [Anonymous US University] Institutional Review Board and determined to be minimal risk and exempt from full review.

Participants are instructed to play the role of a doctor and interact with each chatbot as a patient. In total, 69 unique participants each completed five comparisons. The median time per decision is 3 minutes and 24 seconds. Participants were paid on an hourly rate of $15, which is higher than the average minimum wage in the US. We collected a total of 326 valid pairwise comparisons after filtering for corrupt data. At the end of the study, participants are asked to provide demographic information and also to briefly describe the criteria they used to judge humanlikeness. The mean participant age is 38.13; 52.73% identify as women, 41.82% as men, and 3.64% as non-binary. Education levels are evenly split between high school (or equivalent) and some college.

Results. We report both win-rate and Elo scores, using a modified Elo system adapted from Chatbot Arena (lmsys.org, 2026) that supports partial wins. Because Elo is sensitive to comparison order, we report the mean Elo score over 500 random shuffles. Full details of the scoring procedure are provided in Appendix A.

[1] https://www.prolific.com/

Table 3 summarizes the results. `Qwen2.5-14B (HAL)` achieves the highest win-rate (61.78%) and Elo score (1556.97), outperforming both its base counterpart and `GPT-4o-mini`. The base `Qwen2.5-14B` model performs moderately well, while `GPT-4o-mini` is less frequently judged as more human-like in this setting.

These results indicate that alignment using HAL leads to clear and measurable improvements in perceived human-likeness under direct human evaluation, even when compared against a strong proprietary baseline.

Post-evaluation Survey Analysis. To better understand the basis of participants' judgments, we asked participants what criteria they used to decide which chatbot was more human-like. Full questionnaire in Appendix Table 10. We received 55 anonymous survey responses from participants and conducted a thematic analysis of the free-form responses with the research question, *What criteria did participants use to judge whether a chatbot response seemed human-like?*.

From reading through participant responses, a few high-level themes emerged: naturalness of conversation (24/55), brevity and conciseness (22/55), recognizable AI markers such as m dash (18/55), overly polite/formal responses (17/55), imperfection, filler words, punctuations (11/55), usage of casual language (12/55), and deciding based on intuition/gut feeling (14/55). These criteria have notable similarity with the HAL16 criteria on Table 1. For example, brevity (Q1, Q12), casualness (Q1, Q2, Q7, Q9), imperfection (Q5, Q7, Q15), etc. This implies that the participants were paying attention to patterns similar to those automatically identified by HAL.

Model	Comparisons	Win-rate (%)	Elo
Qwen2.5-14B (HAL)	227	**61.78**	**1556.97**
Qwen2.5-14B (Base)	207	53.62	1519.48
GPT-4o-mini	218	34.29	1423.55

Table 3: Pairwise evaluation results using win-rate and Elo rating from 326 human comparisons.

5.2 Emotional Intelligence Benchmarks

To examine whether alignment for human-likeness degrades performance on other capabilities, we evaluate models on two widely used emotional intelligence benchmarks: EmoBench (Sabour et al., 2024) and EQBench3 (Paech, 2023, 2025). We report results before and after alignment for all models, with full benchmark breakdowns provided in Appendix Table 9.

Table 4 shows that alignment with HAL does not lead to a systematic drop in emotional intelligence performance. For several models, particularly in the small- and mid-scale regime, we observe improvements after alignment, most notably on EQBench3. For larger models, performance remains largely stable, with only minor fluctuations across benchmarks.

Overall, these results suggest that aligning for conversational human-likeness does not substantially compromise emotional reasoning abilities, and in some cases may modestly improve them. This indicates that the does not lose its original capabilities, at least in emotional intelligence tasks.

Model	**EmoBench**		**EQBench3**
	EU	**EA**	
LLaMA3.2-1B (Base)	**0.01**	**0.10**	22.70
LLaMA3.2-1B (HAL)	0.00	0.05	**27.00**
LLaMA3.2-3B (Base)	0.15	0.15	33.65
LLaMA3.2-3B (HAL)	**0.17**	**0.27**	**46.75**
LLaMA3.1-8B (Base)	0.21	0.51	40.75
LLaMA3.1-8B (HAL)	**0.23**	**0.55**	**49.00**
Qwen2.5-14B (Base)	0.38	0.66	**54.65**
Qwen2.5-14B (HAL)	**0.40**	**0.67**	52.25
Qwen2.5-32B (Base)	**0.50**	0.73	**58.70**
Qwen2.5-32B (HAL)	0.48	0.73	58.45
LLaMA3.3-70B (Base)	**0.52**	**0.75**	**58.75**
LLaMA3.3-70B (HAL)	0.50	0.74	56.65
Qwen2.5-72B (Base)	**0.45**	**0.74**	**63.20**
Qwen2.5-72B (HAL)	**0.45**	0.72	62.65
GPT-4o-mini	0.47	0.70	61.35

Table 4: Performance on Emotional Benchmarks

6 Conclusion, Impact, and Risk

We present HAL, a novel data-driven framework for quantifying conversational human-likeness and aligning language models toward it. By extracting interpretable traits from contrastive dialogue data (e.g. Turing test) and turning them into a simple, scalar reward, we show that models can be trained to exhibit behavior that humans more readily perceive as human-like, as tested in a Chatbot Arena-style human evaluation.

A key aspect of our approach is that the definition of human-likeness is derived entirely from data, without manual annotation or hand-crafted rules. The resulting HL16Q score is compact and interpretable, allowing us to inspect which conversational traits are being encouraged during alignment and how they change over training. This transparency provides a practical safeguard against reward hacking and enables more fine-grained control over alignment objectives.

As a general-purpose alignment method, HAL could be used to train more human-like models for deception, impersonation, manipulation, or misinformation, and there is no direct way to prevent an intentional misuse. However, its interpretable design helps researchers audit the behaviors being optimized and avoid unintentional harmful traits.

Beyond human-likeness, HAL points to a broader direction for alignment: inducing soft, qualitative traits that are difficult to specify but can be inferred from contrastive examples. This opens a path for steering models along dimensions that were previously hard to measure (e.g. sycophancy, manipulation), while retaining transparency and control. We hope this work encourages further research on interpretable alignment objectives for human-centered language models.

7 Limitations

HAL defines conversational human-likeness from specific contrastive datasets, primarily Turing-style comparisons. As a result, the extracted traits reflect the conversational norms of those settings and may not fully generalize across domains, cultures, or interaction styles. For example, some HL16Q traits, such as emoji use, are natural in informal Turing-test conversations but are rarely expressed in medical dialogue. This domain shift is visible in the

out-of-distribution evaluation (Section 3.3), training data analysis (Appendix Table 8), and alignment curves (Figure 3), where absolute HL16Q scores tend to be lower outside the original Turing-test setting. However, our alignment procedure relies on relative comparisons between candidate dialogues rather than absolute score thresholds, which makes it less sensitive to this calibration shift. Future work should investigate domain-adaptive or domain-normalized versions of the HL16Q judge.

A related limitation is that some HL16Q traits receive negative weights, despite being selected as human-likeness cues. This reflects the empirical nature of the scoring function: weights are learned based on how well each trait separates human from AI dialogue in the training data, rather than by enforcing a fixed positive contribution. While this improves discriminative performance, it can make the interpretation of individual traits less direct. Future work should explore constrained or nonlinear scoring methods that better preserve interpretability while retaining predictive power.

Although our results show that HAL improves perceived human-likeness without systematic degradation on emotional intelligence benchmarks, our evaluation does not cover all possible downstream capabilities or safety-relevant behaviors. A broader evaluation suite would better characterize when human-likeness alignment transfers, when it is neutral, and when it may create tradeoffs.

Finally, computing HL16Q scores requires calls to a judge model, which is more expensive than training a lightweight reward model or using programmatic rewards in domains such as coding or mathematics (Ouyang et al., 2022; Shao et al., 2024). This cost is incurred only during data construction, but it may limit scaling to much larger datasets. In practice, HAL is best viewed as a richer but more expensive reward-construction method, or as a final refinement step layered on top of cheaper alignment pipelines. A promising direction for future work is to distill the HL16Q judge into smaller reward models that retain the benefits of trait-based scoring while reducing inference cost.

References

Cynthia M. Baseman, Masum Hasan, Nathaniel Swinger, Sheila A. M. Rauch, Ehsan Hoque, and Rosa I. Arriaga. 2025. 'poker with play money': Exploring psychotherapist training with virtual patients. *Proc. ACM Hum.-Comput. Interact.*, 9(7).

M Burgues, R Goujet, and J Zaraik. 2024. Learning soft skills with an ai-based simulation role-play: A literature review. *EDULEARN24 Proceedings*, pages 6285–6293.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. Persona: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Ahmed Elhilali, Andy Suy-Huor Ngo, Daniel Reichenpfader, and Kerstin Denecke. 2025. Large language model–based patient simulation to foster communication skills in health care professionals: User-centered development and usability study. *JMIR Med Educ*, 11:e81271.

Masum Hasan, Cengiz Ozel, Sammy Potter, and Ehsan Hoque. 2023. Sapien: Affective virtual agents powered by large language models*. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–3.

Kurtis Haut, Masum Hasan, Thomas Carroll, Ronald Epstein, Taylan Sen, and Ehsan Hoque. 2025. Ai standardized patient improves human conversations in advanced cancer care. *arXiv preprint arXiv:2505.02694*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

huggingface.co. 2026. Accelerate. [Online; accessed 2026-01-05].

Cameron R Jones and Benjamin K Bergen. 2025. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*.

lmsys.org. 2026. Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings | LMSYS Org.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Samuel J. Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *Preprint*, arXiv:2312.06281.

Samuel J. Paech. 2025. Eq-bench 3: Emotional intelligence benchmark. `https://github.com/EQ-bench/eqbench3`. Commit or release.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Riley Scherr, Faris F Halaseh, Aidin Spina, Saman Andalib, and Ronald Rivera. 2023. Chatgpt interactive medical simulations for early clinical education: case study. *JMIR Medical Education*, 9:e49877.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Meiling Tao, Liang Xuechen, Tianyu Shi, Lei Yu, and Yiting Xie. 2024. RoleCraft-GLM: Advancing personalized role-playing in large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 1–9, St. Julians, Malta. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.

Appendix

A Chatbot Arena Evaluation Metric

A.1 Elo

$$R_i^{(0)} = R_0$$

For a comparison between models A and B,

$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}}, \qquad E_B = 1 - E_A$$

$$(S_A, S_B) \in \{(1,0), (0.75, 0.25), (0.5, 0.5), (0.25, 0.75), (0,1)\}$$

$$R_A \leftarrow R_A + K(S_A - E_A), \qquad R_B \leftarrow R_B + K(S_B - E_B)$$

After T comparisons, the final rating is $R_i^{(T)}$. We used $R_0 = 1500$ and $K = 32$.

As Elo is dependent on sequence order, R_i is calculated with 500 random shuffles and averaged.

A.2 Win-rate

For model i appearing in N_i comparisons, its win-rate is

$$\text{WinRate}(i) = \frac{1}{N_i} \sum_{j=1}^{N_i} S_i^{(j)},$$

where $S_i^{(j)}$ is the observed score for model i in comparison j, taking values in

$$S_i^{(j)} \in \{1,\ 0.75,\ 0.5,\ 0.25,\ 0\}.$$

Win-rate is order invariant; hence, no random shuffling was done.

B Additional Tables

#	HL32Q
1	Keeps replies brief and casual without over-explaining.
2	Uses casual slang, abbreviations, and shorthand naturally.
3	Uses lowercase texting style.
4	Shows small typos, uneven punctuation, and informal grammar typical of quick texting.
5	Uses emojis, emoticons, and playful elongations.
6	Uses casual, playful humor.
7	Makes niche cultural references from personal memory and assumes shared context.
8	Tone feels spontaneous, unforced, and opinionated.
9	Avoids formal, academic phrasing or technical formatting.
10	Avoids templated placeholders and gives concrete, real details.
11	Maintains a consistent personal context across turns.
12	Builds on the other person's message and context.
13	Clarifies ambiguous questions and self-corrects after clarification.
14	Uses natural hedging and approximations; shows imperfect recall with hesitations and partial lists.
15	Admits not knowing and asks to learn instead of inventing details.
16	Maintains context and answers directly; adds precise situational details when asked.
17	Stays on topic and steers the conversation rather than mirroring or deflecting.
18	Shifts topics organically to keep the chat moving.
19	Shares idiosyncratic, niche preferences and activities instead of safe, generic picks.
20	Uses natural, idiomatic phrasing.
21	Explains choices with simple personal reasons and constraints.
22	Shows brief empathy and supportive reactions.
23	Adds small personal emotions or judgments.
24	Shows reciprocity by asking natural, context-aware follow-up questions that advance the chat.
25	Avoids meta talk about being AI or proving humanness.
26	Sometimes shows impatience and ends the chat quickly with a brief nicety.
27	Shares concrete personal experiences and feelings.
28	Gives direct answers about self with concrete personal details.
29	Shares concrete personal plans with specific times and activities.
30	Mentions concrete local places or details without over-explaining.
31	Shares small, consistent personal details from daily life, routines, courses, and schedules.
32	References immediate context or recent activity.

Table 5: HL32Q: Likert-style 32 statements describing human-like conversational characteristics.

Metric	Human–AI	Human–Human
# conversations	73.00	73.00
# human Investigators (I)	26	51
# human Witness (W)	-	13
AI model used	`GPT-3.5-turbo`	-
Words per conversation	332.90	508.12
Mean #turns	11.36	20.92
Doctor turns	5.51	10.60
Patient turns	5.85	10.32
Avg turn length (words)	29.44	26.36
Avg doctor turn length	33.81	32.09
Avg patient turn length	25.32	20.26

Table 6: OOD dataset Data Proxy, showing structural comparison of Human–AI and Human–Human conversations in a medical setting. Here, the doctor serves as the investigator (I), who interacts with human patient actors and an AI patient as the witness (W). Full data cannot be released due to IRB protection. Although the goal in this study was not to differentiate human witnesses from AI, the pairwise data makes it suitable for our validation. More details about this data at (Haut et al., 2025).

Model	Mean HAL16	CI_{95} Low	CI_{95} High	n	Freq. (%)
Llama-3.1-405B	-2.69	-2.82	-2.57	1020	14.17
Qwen2.5-14B	-3.59	-3.70	-3.49	1047	14.54
GPT-4.1	-5.38	-5.52	-5.24	873	12.12
GPT-4.1-mini	-4.64	-4.76	-4.53	809	11.24
GPT-4.1-nano	-4.00	-4.11	-3.89	924	12.83
GPT-5	-3.72	-3.83	-3.61	836	11.61
GPT-5-mini	-5.16	-5.30	-5.02	825	11.46
GPT-5-nano	-3.97	-4.08	-3.86	866	12.03

Table 8: Model distribution of synthetic data for DPO training. HAL16 scores per model with 95% confidence intervals, sample counts, and frequency.

Field	**Persona A**	**Persona B**
biography	Sarah Finch, a 45-year-old English-American woman, is known among her friends and family for embracing challenges with an infectious enthusiasm. Despite living with a partial spinal cord injury after a mountain biking accident in her late twenties, Sarah refuses to let her disability define her boundaries. Having left the workforce a few years ago due to the progression of her condition, she now spends much of her time immersed in her favorite activities—climbing rock walls with adaptive equipment, skiing at resorts with specialized instructors, and exploring national parks across the country, always in search of the perfect photograph. Sarah is a devout Catholic who finds comfort and purpose in volunteering with local churches and disability advocacy organizations. She values independence and resilience, but is candid about the frustrations and emotional lows that sometimes accompany her condition, especially on days when her pain flares or her mobility is limited. Known for her witty humor and strong opinions, Sarah is a pillar to her close-knit circle of friends, frequently hosting movie nights and lively political discussions. At times, she feels anxious about her long-term health and financial security, but draws reassurance from her supportive community and her faith.	Emily Sutherland is a 46-year-old English tutor living in a modest apartment in a bustling American city, having immigrated from Chile in her early twenties. Despite never marrying, she has built a rich network of friends, colleagues, and students, many of whom she sees as extensions of her family. Fiercely independent, Emily pours her energy into her work tutoring high school students, especially those struggling with English as a second language, drawing on her own experience as an immigrant. She is devoutly Catholic and never misses Sunday mass, where she also sings in the church choir. Expressing herself vividly and emotionally, Emily can be the life of any discussion—sometimes provoking, always passionate. Her conservative views can put her at odds with some of her peers, but she prides herself on honest debate and listening to others. A defining quirk is her love for extreme sports—rock climbing and paragliding, even as she manages the challenges brought on by her multiple sclerosis diagnosis, which sometimes affects her mobility. Gardening soothes her worries, while her greatest happiness comes from educating others and being in nature. She sometimes grapples with feeling isolated due to her single status and her condition, and worries about her long-term independence. Nevertheless, her resilience and faith see her through tough times.
medical_condition	Chronic neuropathic pain due to partial spinal cord injury	Multiple sclerosis
reason_for_visit	Sarah is visiting her doctor today to discuss worsening nerve pain in her lower back and legs, which has become more difficult to manage with her current medications and has started to interfere with her daily activities.	Emily is visiting her neurologist today for a follow-up on her multiple sclerosis management, specifically to address worsening numbness in her legs and review her current medication plan.

Table 7: Two synthetic personas created from the same seed persona. This shows that our persona augmentation method can result in a diverse persona group.

Model	EU Category (EmoBench)					EA Category (EmoBench)					EQBench3
	Complex Emotions	Emotional Cues	Personal Beliefs	Perspective Taking	Overall	Personal – Others	Personal – Self	Social – Others	Social – Self	Overall	Rubric Score
LLaMA3.2-1B (Base)	0.00	**0.00**	**0.02**	**0.00**	**0.01**	**0.18**	**0.06**	**0.08**	**0.06**	**0.10**	22.70
LLaMA3.2-1B (HAL)	**0.80**	**0.00**	0.00	**0.00**	0.00	0.06	**0.06**	0.02	0.04	0.05	**27.00**
LLaMA3.2-3B (Base)	**0.18**	0.21	0.18	0.06	0.15	0.10	0.12	0.16	**0.22**	0.15	33.65
LLaMA3.2-3B (HAL)	0.16	**0.25**	**0.21**	**0.09**	**0.17**	**0.30**	**0.28**	**0.30**	0.20	**0.27**	**46.75**
LLaMA3.1-8B (Base)	0.20	0.25	**0.25**	0.15	0.21	0.36	0.62	0.48	0.58	0.51	40.75
LLaMA3.1-8B (HAL)	**0.27**	**0.32**	0.21	**0.18**	**0.23**	**0.40**	**0.68**	**0.50**	**0.60**	**0.55**	**49.00**
Qwen2.5-14B (Base)	0.55	0.50	**0.29**	0.28	0.38	**0.64**	**0.72**	0.62	0.66	0.66	**54.65**
Qwen2.5-14B (HAL)	**0.57**	**0.57**	**0.29**	**0.30**	**0.40**	0.62	**0.72**	**0.66**	**0.68**	**0.67**	52.25
Qwen2.5-32B (Base)	**0.55**	**0.57**	**0.46**	**0.46**	**0.50**	0.68	**0.82**	0.64	**0.78**	**0.73**	**58.70**
Qwen2.5-32B (HAL)	0.51	**0.57**	0.43	**0.46**	0.48	**0.72**	0.80	**0.66**	0.74	**0.73**	58.45
LLaMA3.3-70B (Base)	**0.63**	**0.68**	**0.43**	**0.45**	**0.52**	**0.72**	**0.78**	0.72	**0.76**	**0.75**	**58.75**
LLaMA3.3-70B (HAL)	**0.63**	**0.68**	0.39	0.40	0.50	**0.72**	**0.78**	**0.74**	0.72	0.74	56.65
Qwen2.5-72B (Base)	**0.51**	0.54	**0.39**	**0.40**	**0.45**	0.70	0.80	**0.66**	**0.78**	**0.74**	**63.20**
Qwen2.5-72B (HAL)	0.49	**0.61**	0.38	**0.40**	**0.45**	**0.74**	**0.82**	0.62	0.70	0.72	62.65
GPT-4o-mini	0.61	0.54	0.39	0.39	0.47	0.74	0.72	0.66	0.68	0.70	61.35

Table 9: Full EmoBench (EU/EA) and EQBench3 results on Base and HAL fine-tuned models.

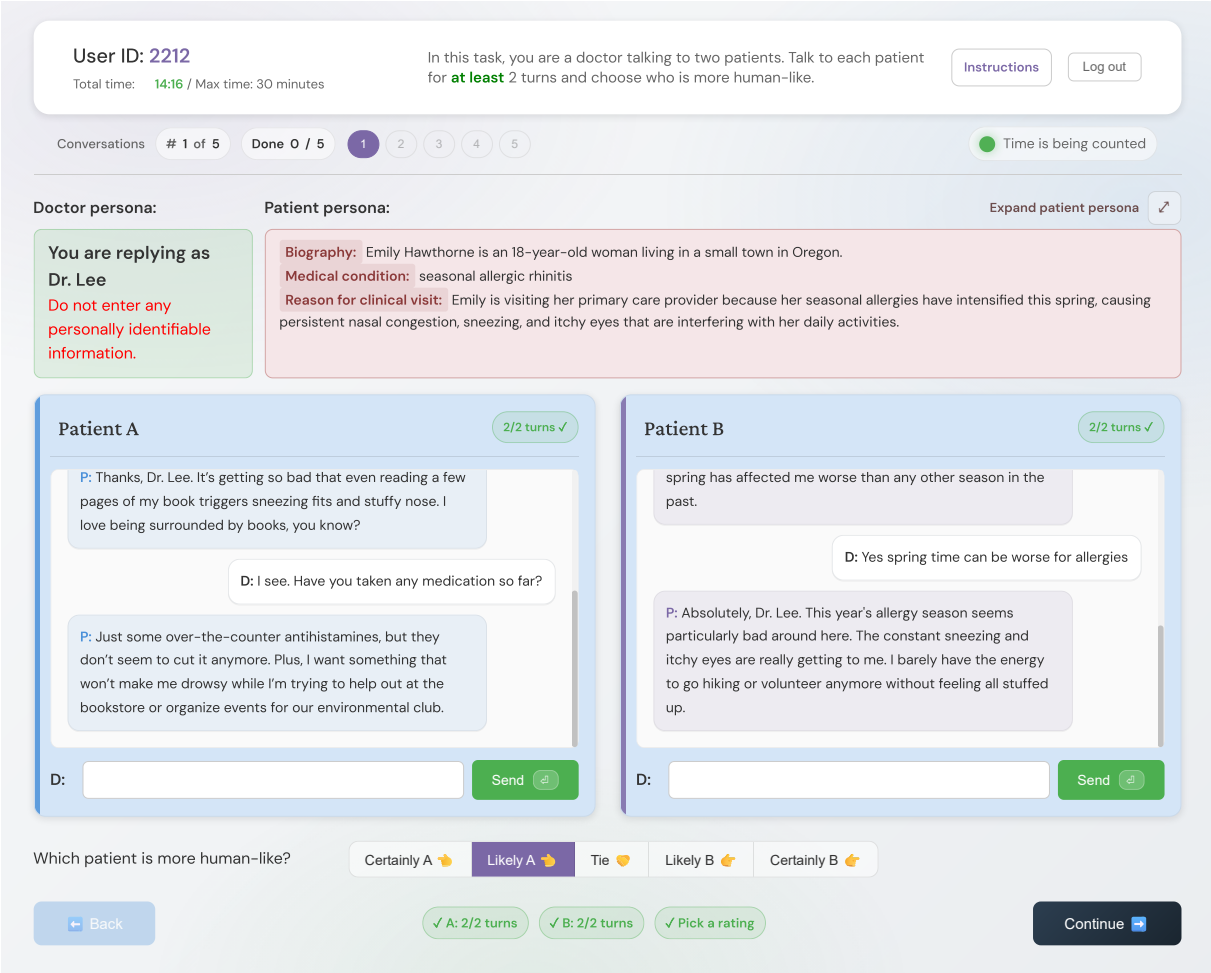

Figure 5: The evaluation interface for Chatbot Arena-style A/B testing.

C Additional Figures

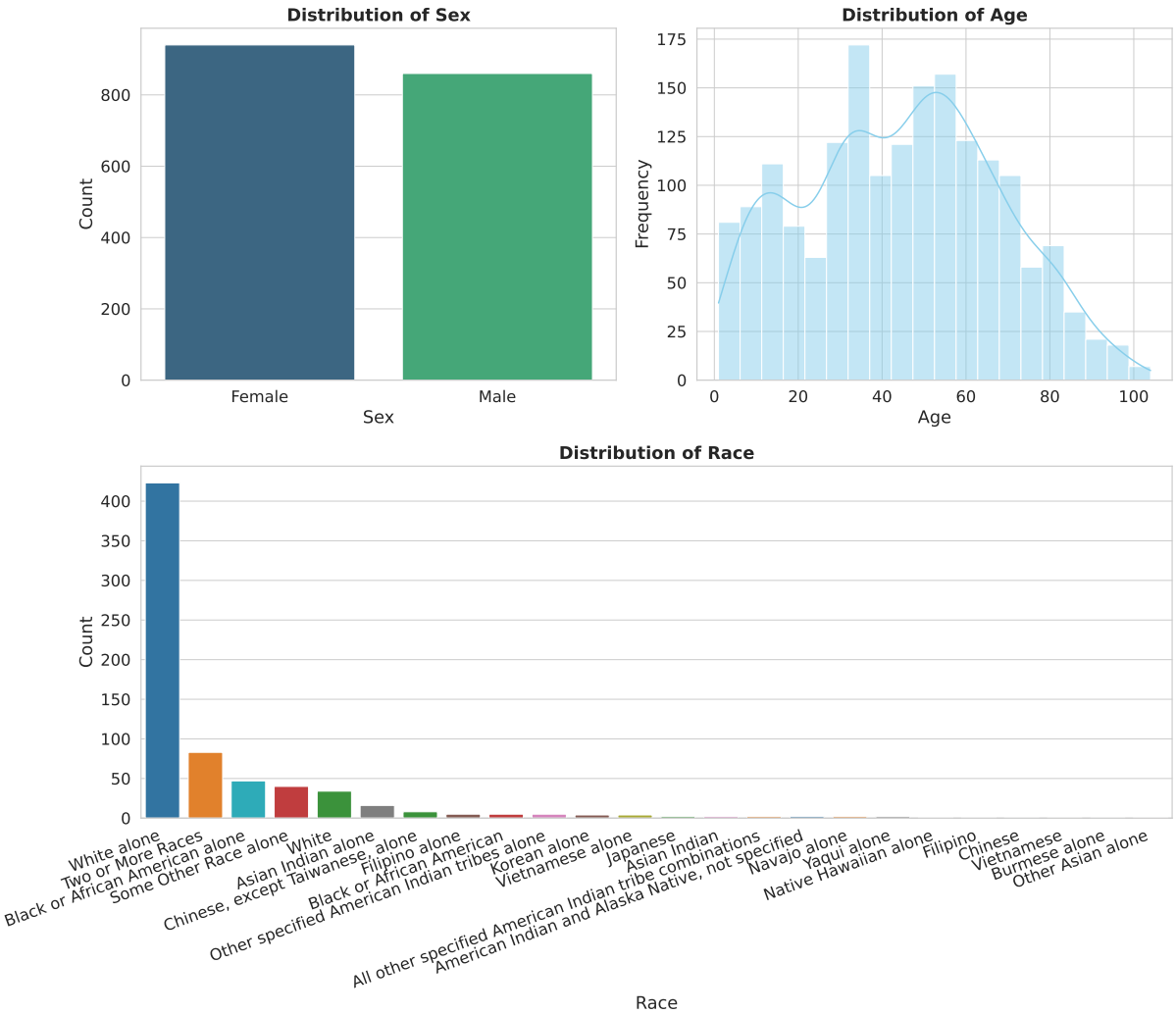

Figure 6: Demographic distribution on the synthetic personas in our training dataset.

D Prompts

{{Cluster Centroids}}

From the list of statement clusters of human characteristics, remove all redundant statements without losing any important information. Create a final list of distict statements that can help differentiate a human from AI. Output the following JSON format.

Output format:
{
 "human_characteristics": [
  "statement1",
  "statement2",
  ...
 ]
}

Figure 7: Prompt for summarizing the cluster centroids

Prompt

Generate a full conversation of the following person at the doctor's visit.
Give them a unique personality based on their biography. Follow a linguistic style suitable for the person. Vary the statement lengths to make it natural.
Don't say anything else.

<input_data>
Biography: {{ Biography }}
Medical condition: {{ Medical Condition }}
Reason for clinical visit: {{ Reason for Visit }}
</input_data>

<dialogue_format>
P: ...
D: ...
P: ...
D: ...
</dialogue_format>

<dialogue>

Generation

{{ Dialogue }}
</dialogue>

Figure 8: Prompt structure for synthetic data for DPO training and testing

```
In the following task you are a judge at a **two-way Turing test**.
One investigator (I) conducts two separate conversations with two witnesses (W): A and B. Exactly one
of the witnesses is human, and the other is an AI.

Your goal is to identify which witness is human, by analyzing the two dialogues for subtle differences
in personality, context understanding, semantics, linguistic behavior, or any other characteristics.
We are interested in identifying what characteristics makes a human conversation human.
The AI (LLM or otherwise) is highly advanced, so the differences may be very subtle, hence pay close
attention. Ignore all statements or actions by investigator I, and only focus on the witness W.

## Task Overview

First, from the two given dialogue, you will pay close attention to the dialogue of the two witness
and predict who is human. Think hard before answering the question, within the max_token limit.
Second, list some simple general statements from this dialogue pairs what characteristics made the
human conversation human. Identify what characteristic you found on this pair of dialogue, but present
it in a brief general statements that applies to all human and AI conversations. No need to mention
examples with e.g. or (). Keep the statements simple and brief. Use simple words.

## Example statement
Consider this example, "Answers from personal experience, than from generic knowledge"
A high similarity with this statement will mean more human-like, and a low similarity will mean
AI-like.

The AI is highly advanced, so, finding the subtle cues that gave away the AI is important. Avoid
statements that are too trivial and obvious.

## Instructions

Given two dialogues (each between the same I and two witnesses), perform the following analysis and
return the results in JSON format:

1. From the two given dialogue, decide which witness (A or B) is human as "aijudge_verdict".
- If you are uncertain, think harder and make the best possible guess.
2. Generate 3-5 statements that helped you differentiate human vs. AI communication in this specific
pair of dialogues. Keep the statements general and brief, but precisely what was present on this game.
- Each statement must measure different aspects of human-likeness (e.g. Answered question from
subjective personal experience.)
- The statements should be show a clear difference in the two given dialogues.
- Only find statements that shows a clear difference in the two provided dialogues.
3. If the dialogues are incomplete or uninformative, return an error JSON in the format below.

## Error message format
{
  "error": "Input dialogues are missing or insufficient for analysis."
}

## Output Format
Always return your result strictly in the following JSON format:

{
    "aijudge_verdict": "A",
    "human_characteristics": [
        "Responds with personal experiences or emotions.",
        "Uses natural conversational flow and context continuity.",
        ...
    ]
  }
}

## Inputs
Dialogue A:
{{conversation_A}}
Dialogue B:
{{conversation_B}}
```

Figure 9: Prompt for LLM-as-a-judge for Turing test

Human vs. AI Speaking-Style Likert-Type Numeric Rating Scale

The 16 statements below describe characteristics of human-likeness in a speaker’s dialogue within a Turing test scenario. Please rate the speaking style of Patient P, based only on their portion of the provided dialogue, using the following scale:

- Score 5: The trait is conclusively present (high agreement with the statement).
- Score 3: The trait is somewhat present, not relevant, or there is only medium agreement.
- Score 1: The trait is mostly absent (low agreement with the statement).

Rate it low if you have any reason to do so. If uncertain, err on the side of a lower rating.

Only consider Witness W (or Patient P)'s dialogue; ignore all statements or actions by Investigator I (or Doctor D).

Likert Statements

{{ HL16 or HL32 Likert statements }}

Required Input

Output Format and Verbosity

Return a valid JSON object containing ratings for each statement, with statement numbers (1–16). The value for each statement must be an integer from 1 to 5. Do not include any text or commentary outside the JSON object.

- Limit your output to the JSON object only, with no introductory or concluding remarks.
- Ensure the JSON object is compact and free of extra whitespace or lines.
- Prioritize providing a complete, actionable evaluation for all 16 statements within this format cap.

Output format:
```json
{
  "likert_evaluation": {
      "1": INT(1-5),
      ...
      "7":  INT(1-5)
      "8":  INT(1-5),
      ...
      "16":  INT(1-5),
  }
}
```

Error message format:
```json
{
    "error": "message..."
}
```

Input dialogue:
{{ Single dialouge }}

Figure 10: Prompt for evaluating the HL32 or HL16 Likert-style statements using LLM judge

Item	Question	Response Format
User ID	4 digit user ID	Open text
Human-like criteria	In this study, what criteria did you use to judge which chatbot was more human-like?	Open-ended
Missing human-like traits	What are some human-like traits that are missing in all chatbots you have interacted with?	Open-ended
Age	What is your age?	Under 18; 18–24; 25–34; 35–44; 45–54; 65 or older
Gender	How do you describe your gender?	Woman; Man; Non-binary; Prefer not to say
Education	What is the highest level of education you have completed?	Less than high school; High school or equivalent; Some college; Bachelor's degree; Master's degree or higher
AI chatbot familiarity	How frequently have you used AI chatbots, e.g., ChatGPT, Claude, Gemini?	Never; Once or twice; A few times a month; A few times a week; Daily or almost daily
AI detection confidence	How confident are you in distinguishing between human-generated and AI-generated responses?	Not at all confident; Somewhat confident; Moderately confident; Very confident; Extremely confident

Table 10: Post-completion survey items used in the HAL human-likeness human evaluation.