HAL: Inducing Human-likeness in LLMs with Alignment

Masum Hasan, Junjie Zhao, Ehsan Hoque
Department of Computer Science, University of Rochester

Paper Link

TL;DR

Conversational human-likeness plays a central role in human-AI interaction: it can help AI understand and respond to us in more natural ways, make AI-based education more effective, or help role-play difficult conversations, such as doctor-patient interactions.

Yet human-likeness remains difficult to **define, measure, and optimize**. In this paper, we propose novel ways to answer these unexplored questions:

1. **What Makes Human Conversation Human?**
2. **How to Induce Those Traits in AI?**

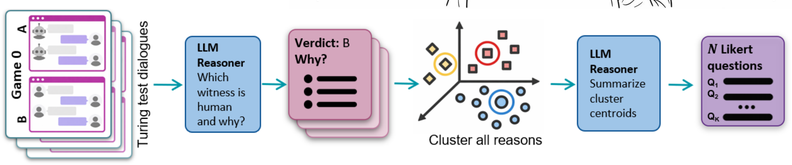
We propose a novel framework for training LLMs named HAL that,

- ✓ Identifies recurring **conversational traits** that reliably distinguish human–human from human–AI dialogue
- ✓ **Compresses** these traits into an interpretable, scalar measure of conversational human-likeness
- ✓ Uses this measure as a **reward signal** for aligning LLMs with standard preference optimization methods
- ✓ Demonstrates through **human evaluation** that model aligned with HAL are more frequently perceived as human-like.

What Makes a Conversation Human?

Turing test: a human judge with a human and an AI and determine who is human.

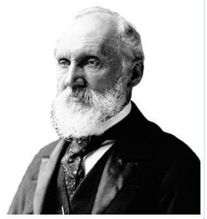
We use a Turing test dataset to find a signal, which traits differ in AI vs human conversation.



- ❖ First, we pass a pair of human-human and human-AI dialogue to an LLM judge, and ask it to identify **who is human and why?**
- ❖ The best model (GPT-5) achieves **64.81% accuracy**.
- ❖ We find the **most common "reasons"** by clustering the text and extracting the centroid.
- ❖ By summarizing the reasons, we find **32 more frequent reasons** (HL32Q) that separate human speaking patterns from AI.

Quantifying Human-likeness

- ❖ To train LLMs to be more human-like, first, we must quantify the abstract trait of "Human-likeness" into a number.
- ❖ We take inspiration from social science and turn the 32 reasons into a Likert survey questionnaire to quantify the unquantifiable.
- ❖ An LLM judge provides a Likert score between 1 and 5 for each of the 32 characteristics in HL32Q for each individual dialogue in the Turing test dataset.
- ❖ Thus, every dialogue becomes a 32-dimensional answer vector (**A**).
- ❖ We train a Linear Regression classifier to learn the weights (**W**) for each dimension and a bias (b).
- ❖ After feature reduction, we are left with 16 statements (**HL16Q**) that have the highest predictability of human-likeness (Table 1).



To measure is to know. If you cannot measure it, you cannot improve it

- Lord Kelvin

Mathematically,

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

Where $N = 32$ and σ(.) is the logistic function. L is the loss function for training logistic regression.

- ❖ By passing a new dialogue with the HL16Q questions with an LLM judge, and multiplying with the associated weights, we can measure the "Human-likeness" of any new dialogue.

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.

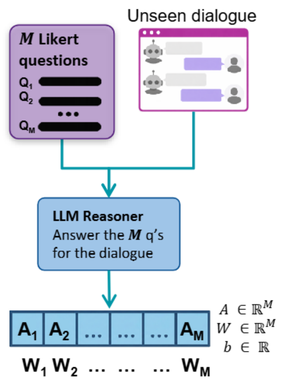


Figure 1: Quantifying human-likeness for an unseen dialogue with HL16Q Judge

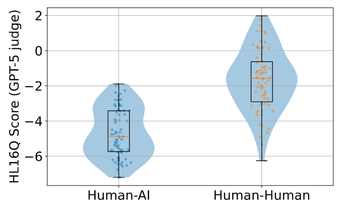
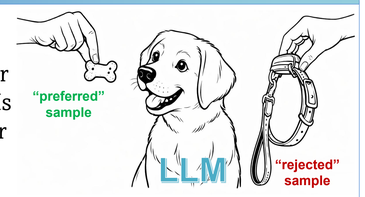


Figure 2: Violin plot of HL16Q Score on Out-of-distribution (OOD) dataset containing Human-AI and Human-Human conversations.

- ❖ We validated HL16Q judge on an out-of-distribution dataset (Figure 2).
- ❖ On a dialogue dataset from a completely different domain, HL16Q Judge statistically significantly separates human-human and human-AI dialogues.

Inducing Human-likeness in LLMs

- ❖ We use **Direct Parameter Optimization (DPO)** – a popular alignment algorithm to train LLMs of 1 billion to 72 billion parameter counts.



- ❖ Alignment algorithms work by showing the model **"preferred"** and **"rejected"** examples and rewarding the model to produce "preferred" work. The same algorithm is used for training LLMs like ChatGPT to produce harmless and helpful responses.
- ❖ We create a **synthetic dataset of 7200 dialogue pairs** of dialogues and evaluate them with HL16Q Judge (GPT-5). Higher-scoring dialogue is considered "preferred", and lower-scoring dialogue is considered a "rejected" sample for training.

Results

Training Performance

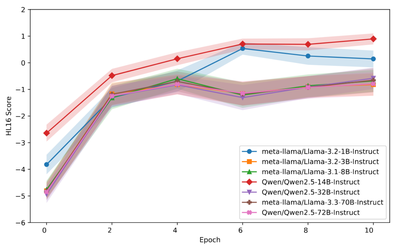


Figure 3: Human-likeness Score increases over 10 Epochs for 1B to 72B parameter model sizes.

Model	EmoBench EU	EmoBench EA	EQBench3
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 3: Models retains its performance EQ benchmarks after training

Human Evaluation

3 LLMs, **326** side-by-side comparisons,
1500+ messages, **64** human participants

HAL is considered more "Human-like"

61.78% times

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

Beyond human-likeness, the HAL framework can induce a range of soft traits in LLMs, aligning them closely with human values.

Contact:
m.hasan@rochester.edu, mehoque@cs.rochester.edu