Does Your AI Agent Get You? A Personalizable Framework for Approximating Human Models from Argumentation-based Dialogue Traces

Supplement

Baselines in Comparative Evaluation

- *HM_1*: An argumentation-based method for updating probability distributions of human models based on argument graphs (Hunter 2015). Inspired by the redistribution function, we apply this concept to our model distribution update.

  At each time step t_i, when an argument A_i is presented by either the agent or the human, we perform the following naive update on the probability distribution:

  $$P_h^{t_i}(m) = \begin{cases} P_h^{t_i-1}(m) + P_h^{t_i-1}(h_{A_i}(m)) & \text{if } m \models A_i \\ 0 & \text{if } m \not\models A_i \end{cases} \tag{1}$$

  where $h_{A_i}(m) = m \backslash \{\alpha\}$ and α is of the form A_i.

- *HM_2*: An enhanced version of Hunter's HM_1 that utilizes the argument structure for updating the distribution (Hunter 2015). Specifically, consider an argument graph G where $\text{Attacks}(G)$ represents the set of attack relations in G. For instance, if $A_1 = \langle \{a\}, \{a\} \rangle$ and $A_2 = \langle \{b, b \to \neg a\}, \{\neg a\} \rangle$, A_2 is a counterargument of A_1, indicating that $(A_2, A_1) \in \text{Attacks}(G)$. In this way, this method first applies Equation (1) and then proceeds with the following update:

  $$P_h^{t_i}(m) = \begin{cases} P_h^{t_i}(m) + P_h^{t_i}(h_\Phi(m)) & \text{if } m \models \Phi \\ 0 & \text{if } m \not\models \Phi \end{cases} \tag{2}$$

  where $\Phi = \{\neg B \mid (B, A_i) \in \text{Attacks}(G) \text{ or } (A_i, B) \in \text{Attacks}(G)\}$.

- *HA*: A state-of-the-art method for learning probability distributions of arguments by Hunter (2016). This baseline method updates the belief in each argument throughout the dialogue by considering the initial probability of each argument and the human's confidence in their arguments. Specifically, for each argument A_i, the final distribution is:

  $$P(A_i) = \begin{cases} 0.2 & x_i = \alpha, \exists B \in \text{Opp}(A_i), P(B) > 0.5 \\ 0.2 & x_i = \eta, \exists B \in \text{Pro}(A_i), P(B) > 0.5 \\ 0.8 & x_i = \alpha, \forall B \in \text{Opp}(A_i), P(B) \leq 0.5 \\ \sigma_i & x_i = \eta, \forall B \in \text{Pro}(A_i), P(B) \leq 0.5 \end{cases} \tag{3}$$

  where $\text{Opp}(A_i) = \{A_{i+1} \mid \exists i, x_{i+1} = \eta\}$ and $\text{Pro}(A_i) = \{A_j \mid \exists j, i < j, x_j = \alpha, (A_j, A_i) \in \text{Attacks}(G)\}$.

	m_1	m_2	m_3	m_4
a	True	True	False	False
b	True	False	True	False
$P_h^{t_0}(m_i)$	0.25	0.25	0.25	0.25
$P_h^{t_1}(m_i)$	0.5	0.5	0	0
$P_h^{t_2}(m_i)$	0	0	1	0

Table 1: An example of the baseline methods.

Example 1. *Suppose there are four possible models, $\mathcal{M} = \{m_1, m_2, m_3, m_4\}$, with a uniform prior distribution $P_h^{t_0}(m_1) = P_h^{t_0}(m_2) = P_h^{t_0}(m_3) = P_h^{t_0}(m_4) = 0.25$. At time step t_1, the agent asserts the argument $A_1 = \langle \{a\}, \{a\} \rangle$. At the next timestep t_2, the human presents the argument $A_2 = \langle \{b, b \to \neg a\}, \{\neg a\} \rangle$ with confidence value $\sigma_2 = 0.6$. Applying the redistribution mechanisms in HM_1 and HM_2 in timesteps t_1 and t_2, respectively, the results are shown in Table 1. By applying HA after timestep t_2, we have $P(A_1) = 0.2$ and $P(A_2) = 0.6$.*

Details of Human Model Approximation

We compared the Spearman's rank correlation distributions in round $k = \{2, 3, 4, 5\}$ of human model rankings where parameters are learned from the previous $k-1$ rounds among Persona and its ablations and the two baselines. Table 2 displays the distribution of Spearman's rank correlation coefficients in each round, showing that Persona performed better than all the other methods in all rounds. Compared to *Generic* and *SBU*, the results demonstrate that incorporating both personalization and the weighting function increases the accuracy of model approximation. Notably, by considering the distribution within $[0.25, 1]$, Persona significantly outperformed HM_1 and HM_2 across all rounds.

We also performed paired Student's t-tests to compare various methods, with Table 3 displaying the p-values that assess the hypothesis that method X outperforms method Y in human model approximation across different rounds. The results show that Persona consistently and statistically significantly outperforms all other methods in almost all rounds, with the exception of Round 5. The reason is that while Persona does better than *Generic* and *SBU* in the high correlation range, they both do better in the medium positive correlation range of $[0.25, 0.75)$ shown in Table 2d. However, the improvements of Persona and its ablation variants *Generic* and *SBU* over state-of-the-art baselines in every round are statistically significant, with p-values smaller than 0.05. These findings demonstrate Persona's ability to leverage existing data to personalize parameters, thereby enhancing human model estimation accuracy in subsequent rounds beyond state-of-the-art baselines and ablation variants. Notably, even the non-personalized ablation variants consistently outperform all baselines, further validating our approach.

Round 2	*Persona*	*Generic*	*SBU*	HM_1	HM_2
$[-1, -0.75)$	0.059	0.065	0.065	0.141	0.152
$[-0.75, -0.25)$	0.147	0.141	0.152	0.185	0.130
$[-0.25, 0.25)$	0.233	0.234	0.207	0.266	0.234
$[0.25, 0.75)$	0.185	0.207	0.250	0.147	0.239
$[0.75, 1]$	0.375	0.353	0.326	0.261	0.245

(a) Comparison of model estimation in Round 2.

Round 3	*Persona*	*Generic*	*SBU*	HM_1	HM_2
$[-1, -0.75)$	0.059	0.076	0.076	0.087	0.081
$[-0.75, -0.25)$	0.114	0.120	0.114	0.179	0.163
$[-0.25, 0.25)$	0.217	0.212	0.234	0.223	0.245
$[0.25, 0.75)$	0.217	0.234	0.223	0.261	0.245
$[0.75, 1]$	0.391	0.359	0.353	0.250	0.266

(b) Comparison of model estimation in Round 3.

Round 4	*Persona*	*Generic*	*SBU*	HM_1	HM_2
$[-1, -0.75)$	0.049	0.054	0.054	0.087	0.071
$[-0.75, -0.25)$	0.130	0.130	0.147	0.158	0.168
$[-0.25, 0.25)$	0.152	0.158	0.152	0.293	0.185
$[0.25, 0.75)$	0.261	0.321	0.266	0.217	0.212
$[0.75, 1]$	0.408	0.337	0.380	0.245	0.364

(c) Comparison of model estimation in Round 4.

Round 5	*Persona*	*Generic*	*SBU*	HM_1	HM_2
$[-1, -0.75)$	0.060	0.060	0.060	0.043	0.065
$[-0.75, -0.25)$	0.076	0.081	0.081	0.136	0.136
$[-0.25, 0.25)$	0.147	0.119	0.130	0.207	0.185
$[0.25, 0.75)$	0.250	0.288	0.277	0.277	0.255
$[0.75, 1]$	0.467	0.451	0.451	0.337	0.359

(d) Comparison of model estimation in Round 5.

Table 2: The distributions of Spearman's rank correlation coefficients in model approximation in Round k ($k = 2, 3, 4, 5$) of human model rankings where parameters are learned from the first $k - 1$ rounds. Note that for participants with only four interactions, the results for Round 5 are identical to those of Round 4.

Post-study Results

In our scenarios, participants were divided into Group A, who ended the conversation themselves, and Group B, where Blitzcrank ended the dialogue. Group A confirmed confidence levels across four rounds, while Group B did so over five rounds. The results in Table 4 provide compelling evidence that the confidence in the AI assistant increases, indicating that participants' confidence grows as the dialogue progresses and the assistant provides more relevant and persuasive arguments.

Finally, Table 5 shows the post-study questionnaire responses further corroborate these findings, with participants reporting high levels of satisfaction with the interaction and the quality of Blitzcrank's arguments.

Round 2 Y X	*Persona*	*Generic*	*SBU*	HM_1	HM_2
Persona	–	0.044	0.047	2.408×10^{-6}	5.730×10^{-5}
Generic	0.956	–	0.187	8.210×10^{-6}	1.907×10^{-4}
SBU	0.953	0.813	–	4.685×10^{-5}	5.659×10^{-4}
HM_1	1	1	1	–	0.908
HM_2	1	1	0.999	0.092	–

(a) The p-values that X outperforms Y in Round 2.

Round 3 Y X	*Persona*	*Generic*	*SBU*	HM_1	HM_2
Persona	–	0.004	0.004	3.254×10^{-4}	0.002
Generic	0.996	–	0.231	0.003	0.021
SBU	0.997	0.769	–	0.005	0.026
HM_1	1	0.997	0.995	–	0.776
HM_2	0.998	0.979	0.974	0.224	–

(b) The p-values that X outperforms Y in Round 3.

Round 4 Y X	*Persona*	*Generic*	*SBU*	HM_1	HM_2
Persona	–	0.010	0.047	3.760×10^{-5}	0.006
Generic	0.990	–	0.748	6.182×10^{-4}	0.030
SBU	0.953	0.252	–	5.905×10^{-4}	0.022
HM_1	1	1	1	–	0.916
HM_2	0.994	0.969	0.978	0.084	–

(c) The p-values that X outperforms Y in Round 4.

Round 5 Y X	*Persona*	*Generic*	*SBU*	HM_1	HM_2
Persona	–	0.640	0.426	0.006	0.001
Generic	0.360	–	0.153	0.003	5.790×10^{-4}
SBU	0.574	0.867	–	0.007	0.001
HM_1	0.994	0.997	0.993	–	0.225
HM_2	0.999	0.999	0.999	0.775	–

(d) The p-values that X outperforms Y in Round 5.

Table 3: The p-values from paired Student's t-tests assessing the hypothesis that X outperforms Y in round k ($k = 2, 3, 4, 5$) in Experiment 2.1

	Group A (Four rounds)	Group B (Five rounds)
$p_{1,2}$	4.199×10^{-15}	1.774×10^{-20}
$p_{2,3}$	3.412×10^{-5}	0.005
$p_{3,4}$	0.02	0.016
$p_{4,5}$	$-$	0.043

Table 4: The p-values of comparing confidence values between interaction rounds. Specifically, $p_{i,j}$ indicates the p-value for the hypothesis that the confidence increases from round i to round j.

	All Participants
Comprehension Score (out of 5)	3.32
Satisfaction Score (out of 5)	3.12

Table 5: Comprehension score and satisfaction score.

References

Hunter, A. 2015. Modelling the Persuadee in Asymmetric Argumentation Dialogues for Persuasion. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Hunter, A. 2016. Persuasion Dialogues via Restricted Interfaces using Probabilistic Argumentation. In *Proceedings of the Scalable Uncertainty Management (SUM)*.