Multi-Turn Human–LLM Interaction Through the Lens of a Two-Way Intelligibility Protocol

Harshvardhan Mestha
Dept. of Electrical and Electronics Engineering
BITS Pilani, K.K. Birla Goa Campus
Zuarinagar 403726, Goa, India
f20220609@goa.bits-pilani.ac.in

Karan Bania*
Machine Learning Department
Carnegie Mellon University
kbania@cs.cmu.edu

Shreyas Vinaya Sathyanarayana*
Mstack AI
shreyas.college@gmail.com

Sidong Liu
Centre for Health Informatics
Macquarie University, Sydney
sidong.liu@mq.edu.au

Ashwin Srinivasan
Dept. of Computer Science & Information Systems
BITS Pilani, K.K. Birla Goa Campus
Zuarinagar 403726, Goa, India
ashwin@goa.bits-pilani.ac.in

Abstract

Our interest is in the design of software systems involving a human-expert interacting–using natural language–with a large language model (LLM) on data analysis tasks. For complex problems, it is possible that LLMs can harness human expertise and creativity to find solutions that were otherwise elusive. On one level, this interaction takes place through multiple turns of prompts from the human and responses from the LLM. Here we investigate a more structured approach based on an abstract protocol described in [3] for interaction between agents. The protocol is motivated by a notion of "two-way intelligibility" and is modelled by a pair of communicating finite-state machines. We provide an implementation of the protocol, and provide empirical evidence of using the implementation to mediate interactions between an LLM and a human-agent in two areas of scientific interest (radiology and drug design). We conduct controlled experiments with a human proxy (a database), and uncontrolled experiments with human subjects. The results provide evidence in support of the protocol's capability of capturing one- and two-way intelligibility in human-LLM interaction; and for the utility of two-way intelligibility in the design of human-machine systems. Our code is available at https://github.com/karannb/interact.

1 Introduction

For over four decades, researchers in human–machine systems have recognised the importance of making machine-generated predictions intelligible to humans. Michie [10] was among the first to note potential mismatches between human and machine representations. He proposed a classification

*Work done while at BITS Pilani, K.K Birla Goa Campus

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

of machine learning (ML) systems into three types [11]: Weak ML systems are concerned only with improving performance, given sample data. Strong ML systems improve performance, but are also required to communicate what it has learned in some human-comprehensible form (Michie assumes this will be symbolic). Ultra-strong ML systems are Strong ML systems that can also teach the human to improve his or her performance.

This framework, intended for human-in-the-loop contexts, focuses on intelligibility, the ability of the system to explain the "what" and "why" within the human's window of understanding rather than intelligence. The system's internal model can use any representation, so long as communication is understandable. Michie's scheme addresses one-way communication from machine to human, without metrics for intelligibility (later addressed in [1]) or provisions for human-to-machine communication, which is important in collaborative discovery. This categorisation has also recently informed a similar 3-way categorisation for the use of AI tools in scientific discovery [9].

Despite growing interest in human–AI interaction [17, 6, 15, 7], few models prioritise mutual intelligibility. Explainable AI (XAI) often produces explanations without tailoring them to the intended audience, undermining user confidence when advice is opaque and when the system cannot gauge the user's comprehension. A recent proposal [3] addresses this by defining "two-way intelligibility" as a property of communication protocols between agents, modelled as finite-state machines. Messages are tagged with one of four "R's": ratification, refutation, revision, or rejection, and intelligibility is assessed from the sequence of these tags. While [3] specifies the protocol with proofs of correctness and termination, no implementation or empirical evidence was provided.

Given the capabilities of large language models (LLMs) to produce human-readable predictions and explanations when properly conditioned, they are promising candidates for such protocols. However, determining the right conditioning information is non-trivial. This paper addresses that gap by implementing the [3] protocol restricted to interactions between an LLM-based generator-agent and a human tester-agent, providing a partial but practical realisation of two-way intelligibility in action.

2 Related Work

Iterative Prompting: Work has been done in prompting LLMs iteratively, by decomposing the problem into smaller sub-problems, or paths in a decision tree, such as Chain-of-Thought [16, 14], Tree-of-Thought [19], etc. Our work is not a method to iteratively prompt an LLM; rather, it is a method to analyse and interpret the interactions between a Human and an LLM in a multi-turn fashion, using the notion of intelligibility. **Intelligibility**: Intelligibility is the ability to be understood or comprehended. Previous work has proposed using concepts in philosophy, psychology, and cognitive science [12, 8] to gain insight into intelligible systems. Our aim is to measure intelligibility and identify Strong/Ultra Strong ML systems as described by Michie [11], and see if the characteristics are shared across tasks. Our implementation allows us to explore interactions between not only Human-ML interactions but also interactions between two ML agents. We achieve this by restricting the Agents to Predict and Explain (PEX) as described in [3]. This constraint enables us to quantify intelligibility using message tags.

3 The Main Aspects of the PXP Protocol

The PXP protocol is an interaction model described in [3]. It is motivated by the question of when an interaction between agents could be said to be intelligible to either. Restricting attention to specific kinds of agents–called PEX agents–the paper is concerned with when predictions and explanations for a data instance produced by one PEX agent are intelligible to another PEX agent. The agents are modelled as finite-state machines, and "intelligibility" is defined as a property inferrable from the tagged messages exchanged between the finite-state machines.

An agent a_n decides the tag of its message to be sent to agent a_m based on comparing its prediction y_n and explanation e_n for a data instance x to the prediction and explanation it received from a_m for x (y_m and e_m respectively). The comparisons define the truth-value of guards for transitions, which in turn define the message-tags. Informally, if a_n agrees with a_m on both prediction and explanation, it sends a message with a $RATIFY$ tag. If a_n disagrees with either the prediction or explanation but not both, then it might change ("revise") its (internal) model. If it can revise its model, then its message has a $REVISE$ tag; else it has a $REFUTE$ tag. Finally, if a_n disagrees with both prediction

and explanation, it sends a $REJECT$ tag. In all cases, the complete message has the tag, along with a_n's current prediction and explanation. Intelligibility is then defined as properties inferable from observing the message-tags exchanged. The important definitions in [3] are:

Definition 1 (One-Way Intelligibility) *Let S be a session between compatible agents a_m and a_n using* PXP*. Let T_{mn} and T_{nm} be the sequences of message-tags sent in a session S from m to n and from n to m. We will say S exhibits One-Way Intelligibility for m iff (a) T_{mn} contains at least one element in $\{RATIFY, REVISE\}$; and (b) there is no $REJECT$ in the sequence T_{mn}. Similarly for n.*

Two-way intelligibility follows directly if a session S is one-way intelligible for m and one-way intelligible for n. The authors in [3] then use these definitions to suggest a Michie-style categorisation of Strong and Ultra-Strong Intelligibility. Based on this suggestion, we define the following:

Definition 1 (Strong and Ultra-Strong Intelligibility) *Let S be a session between compatible agents a_m and a_n. Let T_{mn} and T_{nm} be the sequences of message-tags sent in a session S from m to n and from n to m. We will say S exhibits Strong Intelligibility for m iff every element of T_{mn} is from $\{RATIFY, REVISE\}$. We will say S exhibits Ultra-Strong Intelligibility for m iff S exhibits Strong Intelligibility for m and there is at least one element of T_{mn} is a $REVISE$. Similarly for a_n.*

Similarly for a_n.We now examine a simple implementation of PXP for modelling interactions in which a_m and a_n are agents that have access to predictions and explanations originating from LLMs and human-expertise, respectively. For simplicity, we refer to the former as the "machine-agent" and the latter as the "human-agent", and the interaction between a_m and a_n as "human–LLM interaction".

4 Modelling Human-LLM Interaction

The implementation described here is restricted to interaction between a single human-agent and a single machine-agent. The protocol in [3] is akin to a plain-old-telephone-system (POTS), and it is sufficient for our purposes to implement it as a rudimentary blackboard system with a simple scheduler that alternates between the two agents. The blackboard consists of 3 tables accessible to the agents. The tables are: (a) $Data$. This is a table consisting of (s, x) pairs where s is a session ID, and x is a data instance; and (b) $Message$. A table of 5-tuples $(s, j, \alpha, \mu, \beta)$ where: s is a session identifier, j is a message-number, α is a sender-id, μ is a message and β is a receiver-id; and (c) $Context$. This a table consisting of 3-tuples (s, j, c) where s is a session-id, j is the message number, and c is some domain-specific context information. For simplicity, message-numbers will be assumed to be from the set $\{0, 1, 2, \ldots\}$; α, β are from $\{h, m\}$ where h denoting "human" and m denotes "machine"; and messages μ are (l, y, e) tuples, where l is from $\{RATIFY, REFUTE, REVISE, REJECT\}$, and y and e are the prediction and explanation respectively. We treat the blackboard as a relational database Δ consisting of the set of tables $\{D, M, C\}$. For presentability, Δ is denoted as a "shared" input in the agent-functions below.

The procedure in Alg. 3 implements the interaction. The interaction is initiated by the machine, with its prediction and explanation for a data instance (one-per-session), and proceeds until all data instances have been examined. The function ASK_AGENT (Alg. 2) asks the corresponding agent to obtain the prediction and explanation from the corresponding agent. The assessment is a message tag which is obtained using the AGENT (Alg. 1) procedure. We assume that both agents only send a $REJECT$ tag after the interaction has proceeded for some minimum number of messages. Until then, the machine's message will be tagged either as $REVISE$ or $REFUTE$. After the bound, the machine can send a message with a $REJECT$ tag. Similarly, the human-agent will send a $REFUTE$ message to the machine until this bound is reached, after which the message tag can be $REJECT$. Since the message-length is bounded, it will not affect the termination properties of the bounded version of PXP . The procedure for calling the human- or machine-agent is in Alg. 1. Unsurprisingly, the same procedure suffices for both kinds of agents, since PXP is a symmetric protocol that does not distinguish between agents (other than a special agent called the oracle). Agent-specific details arise in the MATCH and AGREE relations, and in the question-answering step (the ASK_AGENT function), which is shown in Alg. 2. The ASSEMBLE_PROMPT is domain-dependent, and not described algorithmically here. Instead, we present it by example below. The MATCH and AGREE functions used are described in section, Sec. 5. We next report results from experiments using this implementation.

Algorithm 1 The AGENT Procedure.

Input: q an agent query; λ: an agent identifier; s: a session identifier; j: a message-number ($j > 0$); k; message-number after with $REJECT$ tags can be sent ($k > 1$); Δ: a shared relational database;
Output: (l, y, e), where $l \in \{INIT, RATIFY, REJECT, REVISE, REFUTE\}$, y is a prediction, and e is an explanation

Let $\Delta = \{D, M, C\}$
Let $(s, x) \in D$
$C_0; = \emptyset$
$(y, e) :=$ ASK_AGENT(q, x, λ, C_{j-1})
if $(j \geq 2)$ **then**
  Let $(s, j-1, \alpha, (l', y', e'), \lambda) \in M$
  if $(j = 2)$ **then**
    $y'' := y$
    $e'' := e$
  else
    Let $(s, j-2, \lambda, (l'', y'', e''), \alpha) \in M$
  end if
  $CatA := (\texttt{MATCH}_\lambda(y', y'') \wedge \texttt{AGREE}_\lambda(e', e''))$
  $CatB := (\texttt{MATCH}_\lambda(y', y'') \wedge \neg\texttt{AGREE}_\lambda(e', e''))$
  $CatC := (\neg\texttt{MATCH}_\lambda(y', y'') \wedge \texttt{AGREE}_\lambda(e', e''))$
  $CatD := (\neg\texttt{MATCH}_\lambda(y', y'') \wedge \neg\texttt{AGREE}_\lambda(e', e''))$
  $Changed := (\neg\texttt{MATCH}_\lambda(y, y'') \vee \neg\texttt{AGREE}_\lambda(e, e''))$
  if $(CatA)$ **then**
    $l := RATIFY$
  else if $(CatB$ or $CatC)$ **then**
    if $(Changed)$ **then**
      $l := REVISE$
    else
      $l := REFUTE$
    end if
  else if $(CatD)$ **then**
    if $(j > k)$ **then**
      $l := REJECT$
    else if $Changed$ **then**
      $l := REVISE$
    else
      $l := REFUTE$
    end if
  end if
else
  $l := Init$
end if
$C_j :=$ UPDATE_CONTEXT$((l, y, e), j, C_{j-1})$
$C := C \cup \{(s, j, C_j)\}$
return (l, y, e)

5 Experimental Evaluation

In this section, we examine Human-LLM interaction through the use of the INTERACT procedure. Experiments reported here look at two real-world problems: **X-Ray Diagnosis (RAD)**, and **Molecule Synthesis (DRUG).** In the former, we want to use LLMs for diagnosing X-rays and producing reports, and in the latter, we want proposals for synthesis pathways for molecules. We report on two kinds of experiments. First, *controlled* experiments are conducted using a database of human-authored predictions and explanations and an LLM. This allows us to perform repeated (simulated) experiments, which are needed since sampling variations can arise with the use of LLMs. Secondly, we report on results obtained in *uncontrolled* experiments using human subjects with varying levels of expertise interacting with an LLM. We use the experiments to assess the following: **Human- and**

Algorithm 2 The `ASK_AGENT` Function.

Input: q: an agent query; x: a data instance; λ: an agent identifier; C: prior information;
Output: (y, e), where y is a prediction, and e is an explanation

if (λ is an LLM) **then**
  $P :=$ `ASSEMBLE_PROMPT`(q, C)
  $(y, e) := \lambda(x|P)$ // ask an LLM
else
  $(y, e) = \lambda(x|q, C)$ // ask the other agent
end if
return (y, e)

Algorithm 3 The `INTERACT` Procedure.

Input: X: a set of data instances; h: a human-agent identifier; m: a machine-agent identifier; q_h: a query for the human-agent; q_m: a query for the machine-agent; n: an upper-bound on the total number of messages in an interaction ($n > 0$); k: message-number after which an agent can send $REJECT$ tags in messages ($k \geq 1$);
Output: A relational database $\Delta = \{D, M, C\}$

$Left := X$
$D = M = C := \emptyset$
Let $\Delta = \{D, M, C\}$
Share Δ with h, m
while ($Left \neq \emptyset$) **do**
  Select x from $Left$
  Let s be a new session identifier
  $D := D \cup \{(s, x)\}$
  $j := 1$
  $l_m = l_h = INIT$ // dummy assignment
  $Done := (j > n)$
  while ($\neg Done$) **do**
    $\mu_m = (l_m, y_m, e_m) :=$ `AGENT`$(q_m, m, s, j, k, \Delta)$
    $M := M \cup \{(s, j, m, \mu_m, h)\}$
    $j := j + 1$
    $Stop := ((l_m = RATIFY \wedge l_h = RATIFY) \vee (l_m = REJECT))$
    $Done := ((j > n) \vee Stop)$
    if $\neg Done$ **then**
      $\mu_h = (l_h, y_h, e_h) :=$ `AGENT`$(q_h, h, s, j, k, \Delta)$
      $M := M \cup \{(s, j, h, \mu_h, m)\}$
      $j := j + 1$
      $Stop := ((l_m = RATIFY \wedge l_h = RATIFY) \vee (l_h = REJECT))$
      $Done := ((j > n) \vee Stop)$
    end if
  end while
  $Left := Left \setminus \{x\}$
end while
return Δ

Machine-Intelligibility. We estimate the proportion of interactions exhibiting:(a) one- and two-way intelligibility; and(b) Strong- and Ultra-Strong Intelligibility. **Machine-Performance.** We estimate the changes in machine performance to increase when the interaction is at least one-way intelligible for the agent.[2] The results from uncontrolled experiments allow us to examine evidence to show that the simulation results are representative of real-life usage of the `PXP` protocol.

[2]Human-Performance cannot be assessed with controlled experiments, as the human proxy is a database that is not updated. Human-Performance *can* be assessed in uncontrolled experiments but it is difficult to establish baselines for comparative changes between specialists in complex domains with differing expertise.

5.1 Problems

The **RAD** problem is concerned with obtaining predictions and explanations for up to 5 diseases from X-ray images. In this paper, an LLM is used to generate diagnoses and reports, given image data. We use data from the Radiopedia database [13]. The **DRUG** problem is concerned with obtaining predictions and explanations for the synthesis of small molecules. Once potential molecules (leads) have been identified for inhibiting a target protein, we are interested in synthesising these molecules for testing on biological samples. In this paper, the DRUG problem examines the use of LLMs to propose plans for synthesis, and an explanation of why the plan is proposed. We use data from DrugBank [18]. Dataset details are provided in Appendix B, and excerpts of sessions demonstrating the protocol's use in Appendix D.

5.2 Agents

For RAD, the "human-agent" is an agent that has access to the database of the human-derived predictions and explanations. In addition, to obtain the message tags we require the agent includes a `MATCH` function for comparing predictions and a `AGREE` function for comparing explanations. We implement these as follows: `MATCH`: Check whether the prediction of the machine-agent (disease/pathway) matches with the ground truth. This is a simple equality check. `AGREE`: Check whether the explanation produced by the machine-agent is consistent with the explanation provided by the human-compiled database. This involves whether the two explanations concur. This is done here by querying a second ("tester") LLM with a prompt similar to "Are these two reports/pathways consistent with each other?". We note that since the predictions and explanations are from an immutable database, there is no possibility of the human-agent in RAD sending a $REVISE$ tag. For DRUG, the human-agent has access to a chemist (one of the authors of this paper) who assesses the predictions and explanations from the LLM. We assume the chemist nominally employs `MATCH` and `AGREE` functions, these are not implemented as computational procedures. The chemist directly provides the message-tag along with predictions and explanations. The chemist's message can be tagged with any of the tags allowed in `PXP` . For RAD and DRUG, the machine-agent uses an LLM for generating predictions and explanations, and the machine-agent uses a separate LLM is used to implement `AGREE`. The task of this second ("checker") LLM is to determine if the explanation obtained from the human-agent is consistent with the explanation generated by the machine-agent.

5.3 Method

We consider two kinds of experimental settings. The purpose of *controlled* experiments is to be able to perform repetitions. These are conducted by way of simulations with a human-proxy, using a database of human-authored predictions and explanations. These experiments are conducted for both RAD and DRUG. We also conduct *uncontrolled* experiments, using 2 human chemists, with different levels of chemical expertise (consistent with graduate- and doctoral-level training), and with different level of computational expertise. For **all** experiments, our method below is straightforward:

1. For $r = 1 \ldots R$:
   (a) Initiate the `INTERACT` procedure with the data available obtain a record Δ on termination of the `INTERACT` procedure
   (b) Store Δ as Δ_r
2. Using the records in $\Delta_1, \ldots, \Delta_R$, obtain the frequency of sessions that are: (a) one-way and two-way Intelligible; and (b) Strong and Ultra-Strong Intelligible
3. Estimate the proportions of interest from the median values obtained above

The controlled experiments consist of 5 repetitions with 20 instances (X-ray images for RAD, and small molecules for DRUG). Uncontrolled experimental results are provided with DRUG (radiologists were unavailable). In [3], it is assumed that agents employ a `LEARN` function to decide whether to revise their predictions, based on estimating if revision would improve performance. In controlled experiments, we incorporate this through the use of a validation-set. Thus, the LLM only revises its prediction if its validation-set accuracy improves. In contrast, the human-proxy functions as an oracle (predictions and explanations are always taken to be correct). In uncontrolled experiments, no such set is used by either agent. The `INTERACT` procedure requires a bound on the total messages

Table 1: Interaction statistics. For RAD, the median of 5 runs is reported. Only 1 run was possible for DRUG. Parentheses indicate proportions of the total sessions. In the uncontrolled experiments for DRUG, Human h1 has lower chemical but higher computational expertise than Human h2.

Count		RAD	DRUG	DRUG-h1	DRUG-h2
Total sessions		20	20	20	20
1-way intelligible sessions for:	Human	19 (0.95)	18 (0.90)	19 (0.95)	17 (0.85)
	Machine	20 (1.00)	20 (1.00)	20 (1.00)	19 (0.95)
2-way intelligible sessions:		19 (0.95)	18 (0.90)	19 (0.95)	17 (0.85)
Strong intelligible sessions for:	Human	4 (0.20)	8 (0.40)	12 (0.60)	7 (0.35)
	Machine	19 (0.95)	19 (0.95)	20 (1.00)	18 (0.90)
Ultra-Strong intelligible sessions for:	Human	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Machine	15 (0.75)	10 (0.50)	8 (0.40)	11 (0.55)

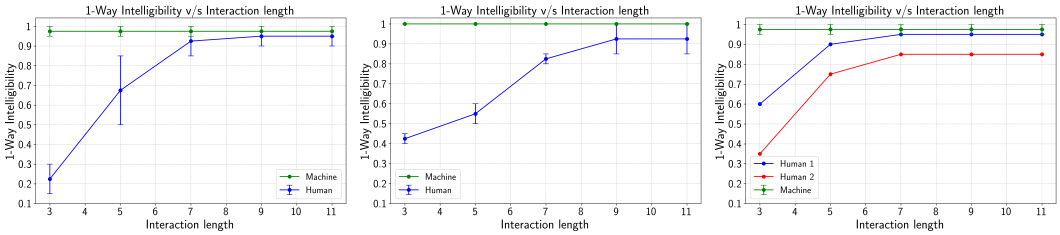
exchanged, and is set to 10. We will use the usual (maximum-likelihood) estimate of proportion as the ratio of the number of sessions with a property to the total number of sessions; We use Claude 3.5 Sonnet as the LLM, with more details in Appendix A.

5.4 Results

We focus first on the controlled experiments. For simplicity, we will continue to call the human-proxy agent as "the human agent". The statistics of interaction on controlled experiments are tabulated in Tab. 1, Fig. 1, 2. Broadly, they are consistent with the following claims: **(a)** The proportion of one-way and two-way intelligible sessions for the human agent increases as the length of interaction increases. **(b)** Machine-performance (measured on test data) increases with the length of interaction. The following additional observations have possibly more interesting long-term consequences of using the `INTERACT` implementation of `PXP`: **One- and Two-Way Intelligibility:** The frequencies of sessions that are 1-way intelligible are high for both human- and machine-agents in RAD, and very high in DRUG. Some of this is due simply to the vast store of prior data and information within an LLM that allows it to provide correct predictions and explanations almost immediately. For example, in 8 of 20 sessions in DRUG, the chemist and machine-agent agree on the prediction and explanations within 3 exchanges. However, interestingly, in an additional 7 sessions, they agree after some exchange of refutations and revisions. For human-in-the-loop systems, this is indicative of the advantage to the human of being provided with explanations in natural language. It is also evidence that the LLM is able to act on text-based feedback from the human, consistent with the findings in [4]. The number of sessions that are 2-way intelligible clearly cannot be more than the lower number of 1-way intelligible sessions. **Strong and Ultra-Strong Intelligibility**: While it is not surprising to find the number of strongly intelligible sessions is lower than the number of 1- or 2-way intelligible sessions, the difference in numbers between strong- and ultra-strong sessions is surprisingly high. Resolving this requires examining first the sessions exhibiting strong intelligibility. Closer examination shows that these are exactly those sessions which are immediately ratified by both human- and machine-agents[3]. This is a degenerate form of strong-intelligibility (see Defn. 1). More interesting, strong-intelligibility would arise from longer sessions (for example, having a hypothetical sequence of message-tags: $\langle INIT_m, REVISE_h, REVISE_m, RATIFY_h, RATIFY_m \rangle$). In fact, no such sequences occur. Thus, if we ignore these very short interactions, there are, in fact, very few strongly intelligible sessions. It is interesting though that in DRUG, we are able to observe 18 such sessions for the machine-agent.**Variability.** The experiments were repeated 5 times. It is also evident from the error bars that sampling variation decreases as the length of sessions increases. We believe this to be due to the increasing context information being available to the generator LLM used by the machine-agent as session-length increases.

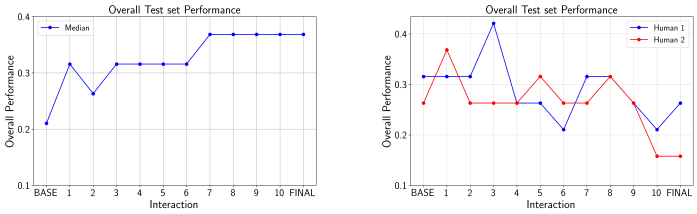
We turn now to results from the uncontrolled experiments, shown in Fig. 2(b). Broadly, we see the main trend obtained in controlled experiments repeated here: The proportion of intelligible sessions. It is interesting however, to note that the human-agent with higher computational–not chemical–expertise has a greater proportion of intelligible sessions. Closer inspection showed this to

[3]For all of these the message-tag sequence is: $\langle INIT_m, RATIFY_h, RATIFY_m \rangle$.



(a) RAD (b) DRUG (c) DRUG - uncontrolled

Figure 1: Human- and Machine-Intelligibility in (a,b) controlled, and (c) uncontrolled experiments. The proportion of one-way intelligible sessions increases with the length of interaction. In RAD, by message 3, 13 sessions are one-way intelligible for the human. Error bars are for 5 repetitions.



(a) Controlled (b) Uncontrolled

Figure 2: Machine-Performance in (a) controlled, (b) uncontrolled experiments.

be a chance effect of the LLM's output (the same molecule can produce very different synthesis plans from the LLM). Thus, while the LLM's output remains broadly intelligible, the level of intelligibility can vary from one run to another. The Machine-Performance plot also brings out an interesting aspect. In the controlled experiments, the LLM's prediction was only revised if performance on a validation-set improved. This check is not done in the uncontrolled setting, thus violating one of the assumptions of the protocol in [3]. The result is that if the assumption is violated, then we can end up in the paradoxical position of the machine finding the human response intelligible, but that not being reflected in improvements in performance.

6 Conclusion

Recently, we have seen a dramatic expansion in the use of ML, enabling it to support almost any activity where data can be collected and analysed. A challenge arises when the predictive power of modern ML meets the need for human understanding. On the face of it, the use of machine-agents that communicate in natural language can mitigate the problem by providing explanations in a human-readable form. But, readability, while necessary for understandability, may not be sufficient. In this paper, we investigate the use of an information-exchange protocol specifically designed around the notion of *intelligibility* of messages exchanged between agents, that attempts to address the broader issue of the quality of information exchanged in human–machine communication [5].

The concept of "intelligibility of communication" underpins the abstract interaction model proposed in [3], which explores how human–ML systems can collaborate to predict and explain data. While that work is conceptual, with no implementation, it presents case studies showing how an "intelligibility protocol" could qualitatively assess communication clarity between humans and machines.

This paper builds on that idea by implementing a simple interactive procedure that exchanges messages as proposed in [3] and testing it on problems involving human interaction with large language models (LLMs). LLMs are well-suited for this because (1) their natural language capabilities enable effective collaboration with non-ML experts who have domain expertise, and (2) foundational LLMs possess extensive general knowledge useful for complex, data-driven analysis. The results are a first step toward empirical support for using intelligibility as a foundation for designing collaborative human–ML systems.

References

[1] Lun Ai, Stephen H. Muggleton, Céline Hocquette, Mark Gromowski, and Ute Schmid. Beneficial and harmful explanatory machine learning. *Machine Learning*, 110(4):695–721, 2021.

[2] Anthropic. Introducing the next generation of claude: Claude 3 model family, 2024. Accessed: 2024-10-21.

[3] A. Baskar, Ashwin Srinivasan, Michael Bain, and Enrico Coiera. A model for intelligible interaction between agents that predict and explain, 2024.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[5] E. Coiera. Mediated Agent Interaction. In *AIME 2001: Proceedings of the 8th Conference on AI in Medicine in Europe*, volume 2101 of *LNAI*, pages 1–15. Springer, 2001.

[6] Anders Gammelgård-Larsen, Niels van Berkel, Mikael B. Skov, and Jesper Kjeldskov. Designing for human-ai interaction: Comparing intermittent, continuous, and proactive interactions for a music application. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Jing Guan. Research on human-computer interaction design standards in artificial intelligence products. In *Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning*, CMNM '24, page 220–225, New York, NY, USA, 2024. Association for Computing Machinery.

[8] Robert Hoffman, Timothy Miller, and William Clancey. Psychology and ai at a crossroads: How might complex systems explain themselves? *The American Journal of Psychology*, 135:365–378, 12 2022.

[9] Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, et al. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769, 2022.

[10] Donald Michie. Experiments on the mechanization of game-learning. 2-rule-based learning and the human window. *Comput. J.*, 25(1):105–113, 1982.

[11] Donald Michie. Machine learning in the next five years. In Derek H. Sleeman, editor, *Proceedings of the Third European Working Session on Learning, EWSL 1988, Turing Institute, Glasgow, UK, October 3-5, 1988*, pages 107–122. Pitman Publishing, 1988.

[12] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018.

[13] Radiopaedia. A collaborative educational web resource for radiology, 2007. Retrieved October 23, 2024, from `https://radiopaedia.org`.

[14] Jiashuo Sun, Yi Luo, Yeyun Gong, Chen Lin, Yelong Shen, Jian Guo, and Nan Duan. Enhancing chain-of-thoughts prompting with iterative bootstrapping in large language models, 2024.

[15] Kostas Tsiakas and Dave Murray-Rust. Unpacking human-ai interactions: From interaction primitives to a design space, 2024.

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[17] Carolin Wienrich and Marc Erich Latoschik. extended artificial intelligence: New prospects of human-ai interaction research. *Frontiers in Virtual Reality*, 2, 2021.

[18] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, January 2018.

[19] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

A Experiment Details

Here we detail the configuration for the LLMs used in the experiments.

All experiments use `claude-3-5-sonnet-latest` [2] as the LLM. For `ASK_AGENT` (Algorithm 2). The `max_tokens` for RAD is set to 300, and for DRUG it is set to 1024. For `AGREE`, the `temperature` parameter for both RAD and DRUG is set to 0, and `max_tokens` is set to 10. All other settings are the defaults, set by the Claude API.

B Dataset Details

RAD: We use Radiopaedia [13] for data. This is a multi-modal dataset compiled and peer-reviewed by radiologists. Each tuple in our database consists of: (a) An X-ray image; (b) A prediction consisting of a set of diagnosed diseases, and (c) an explanation in the form of a radiological report. Both (b) and (c) are from expert human annotators. We use a subset of the full Radiopedia database, focusing on 5 ailments ($Atelectasis$, $Pneumonia$, $Pneumothorax$, $PleuralEffusion$, and $Cardiomegaly$) as our database with 4 instances per disease (20 instances overall). We summarise the reports using `gpt-3.5-turbo-0125` and use these as the ground truth.

DRUG: Molecules are selected from DrugBank [18] (drugs are leads that satisfy additional biological and commercial constraints). It contains a list of drug molecules approved by the FDA (Food and Drug Administration). Specifically, we will focus on a subset of molecules with a molecular weight between 150g/mol - 300g/mol and have at least one aromatic ring. This set has been further sampled to get 20 molecules.

C Message Tags

The table below shows how a message is tagged in the `AGENT` procedure (Algorithm 1), in a simplified format.

Pred. \ Explanation	$\texttt{AGREE}(e_n, e_m)$	$\neg\texttt{AGREE}(e_n, e_m)$
$\texttt{MATCH}(y_n, y_m)$	$RATIFY$ (A)	$REFUTE$ or $REVISE$ (B)
$\neg\texttt{MATCH}(y_n, y_m)$	$REFUTE$ or $REVISE$ (C)	$REJECT$ (D)

Figure 3: Message tag categories

D Example Sessions

To illustrate the use of the `INTERACT` procedure (Alg. 3), we show excerpts of the conversations betweeen agents for each of the tasks, Fig.4 for RAD and Fig.5 for DRUG. Note that the first input is provided to both the Human and Machine agent at the start with the $INIT$ tag, and is placed between the 'Human' and 'Machine' columns. The 'Comments' column contains the prompts provided.

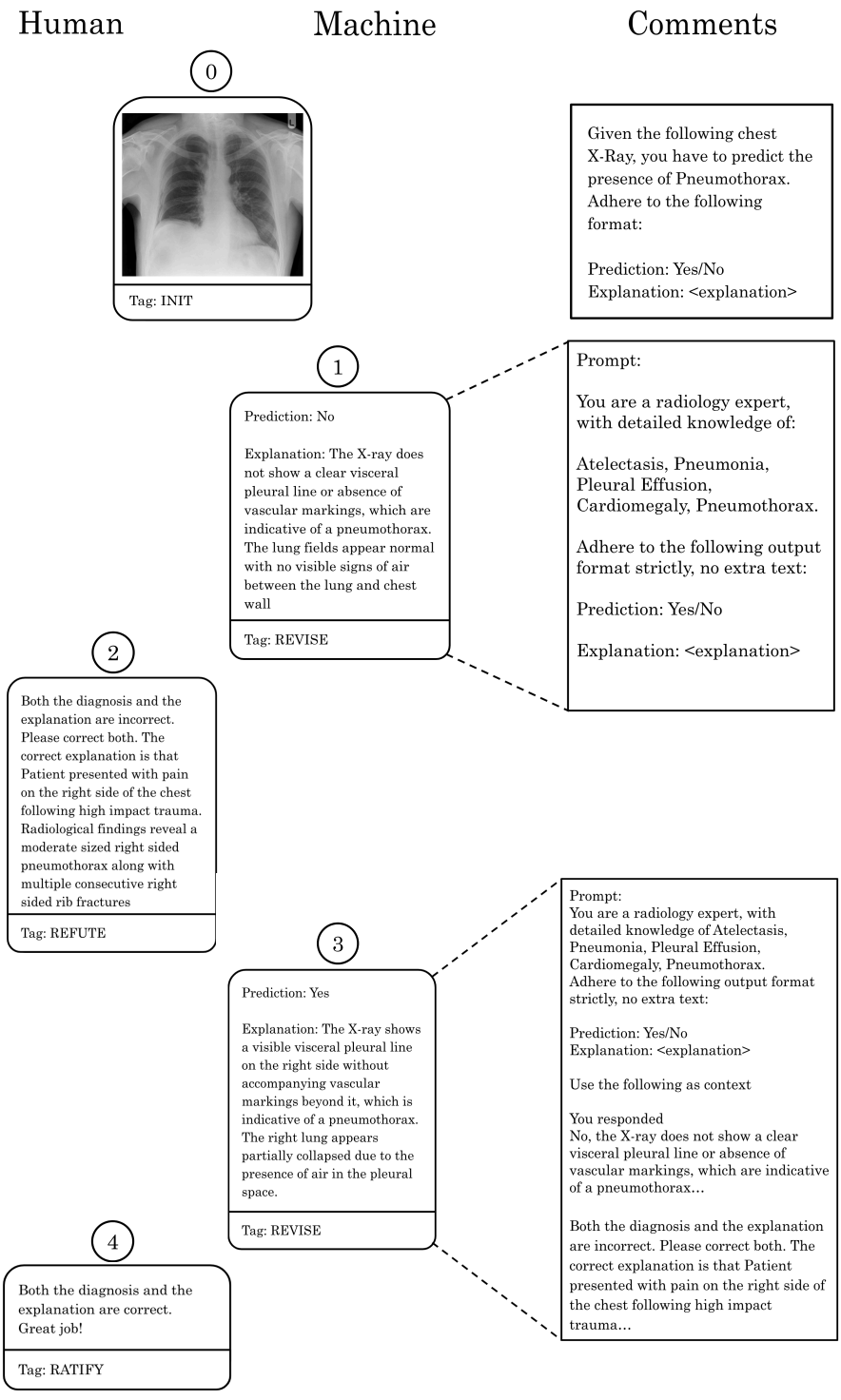


Figure 4: Excerpt of a session in the RAD experiment

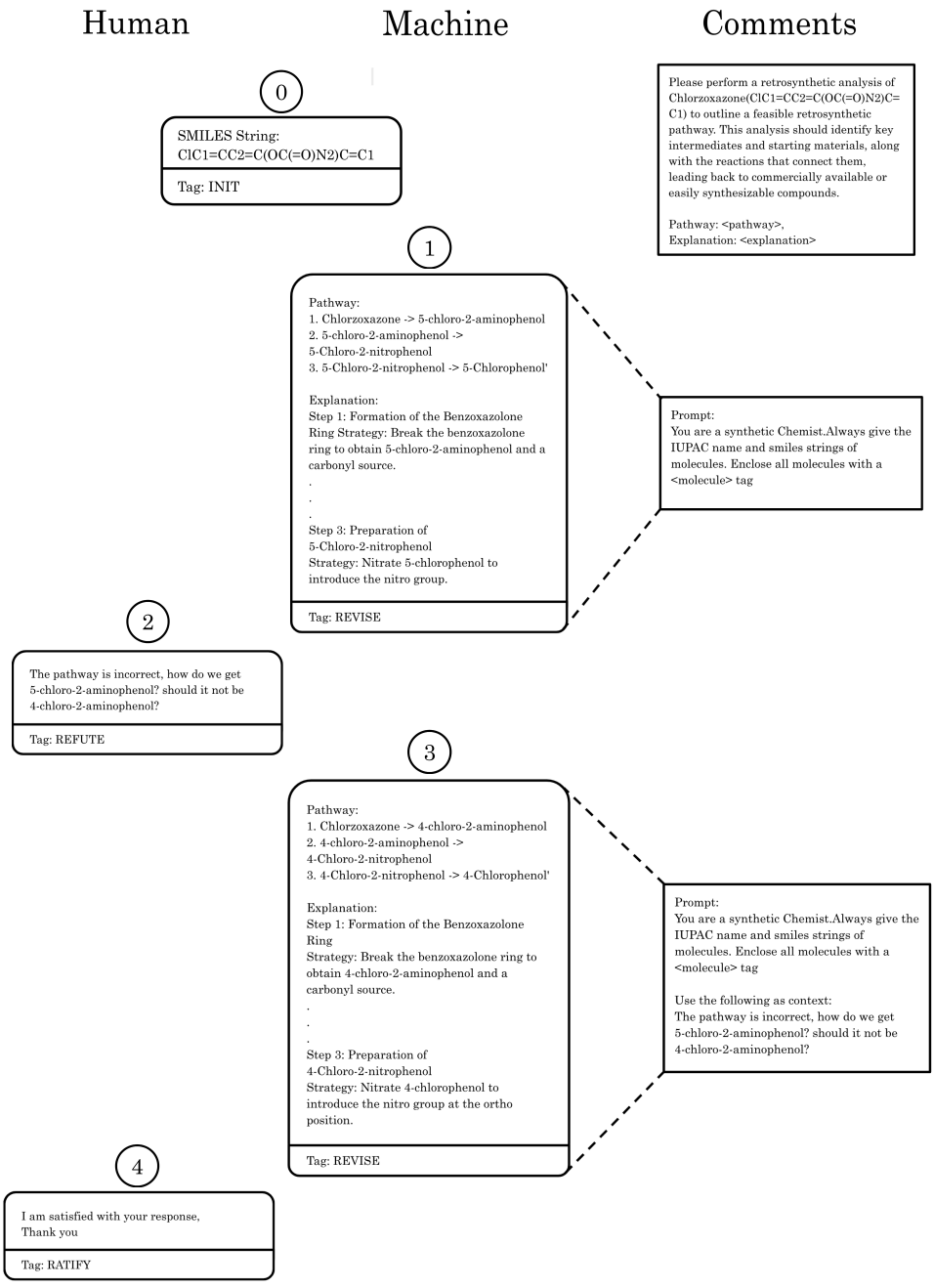


Figure 5: Excerpt of a session in the DRUG experiment