"Why Do You Say So?" Dialogical Classification Explanations in the Wild and Elicited Through Classification Games

Jana Götze and **David Schlangen**
Computational Linguistics, Department of Linguistics
University of Potsdam, Germany
`jana.goetze,david.schlangen@uni-potsdam.de`

Abstract

Enabling classification models to deliver successful explanations requires such models to not just deliver an explanation on top of their classification, but to adapt to the explainee in presenting arguments and details that the explainee may ask about. We present data collection settings that aim at eliciting such dialogical classification explanations in the context of visual dialog where dialog players need to draw conclusions based on this image. We then describe data from a naturally occurring setting as well as two game settings and how the preliminary data we have collected can inform model building.

1 Introduction

Neural network methods have pushed the boundaries of automating classification tasks in many areas of research. For tasks that involve language and image data, such deep learning models are able to reach ever increasing accuracy. A common concern with any kind of automated classification method is its opacity in terms of how it derives its decision based on the training data it was fed. This concern does not just target the human desire to understand what is going on internally but also the necessity to detect and correct possible mistakes or undesired biases in the underlying data. Furthermore, a model's reasoning can inform us about patterns in the data or training setup that can either inform human decision-making or correct misconceptions about cause-effect relations, as case studies from deployed models show (Caruana et al., 2015).

The area of Explainable AI has become more active again in recent years, focusing on different types of methods that can make a model understandable in some way.[1] These methods commonly present a user with an explanation alongside their classification, without the user being able to further question this explanation. In Miller (2019)'s terms, these methods focus on the *cognitive process* of finding a decision rather than the *social process* of delivering the decision to a specific interaction partner.

In many everyday scenarios, human decisions are open to debate: a decision can be questioned or challenged, or a listener may want to ask questions about how the decision came about. For example, students want to understand a grading decision or patients a medical diagnosis. These explanations are user-dependent, i.e. the explainer takes the listener's previous knowledge or intent into account (Miller, 2019).

The structure of such explanatory dialog has been studied by Walton (2009) and has been empirically affirmed by Madumal et al. (2019), who have annotated naturally occurring data of different multi-party explanation settings.

In his survey of sociological research on explanation, Miller (2019) points out that current research on Explainable Artificial Intelligence systems misses some features that human explanation-giving is known to posses from various studies in sociology and neighboring disciplines. One aspect that we take up here is the aspect of interactivity. In conversation, explanations are not given in isolation, but are embedded into the context of a decision that is reached by two or more speakers in a collaborative process and that takes into account the explainee's knowledge about the issue. Research has shown that conclusions that are reached in such a collaborative fashion are more often correct than an average of individual conclusions (Karadzhov et al., 2022b). Agents that can explain their claims on human terms are therefore

---

[1]We acknowledge that there is no consensus about terms such as transparency, explainability, interpretability, justifiability and others. We use the term *explainable* in a loose way that incorporates aspects of any of the other commonly used terminology. We are interested in models that can express their reasoning about a decision in natural language, thus giving a human-understandable explanation.
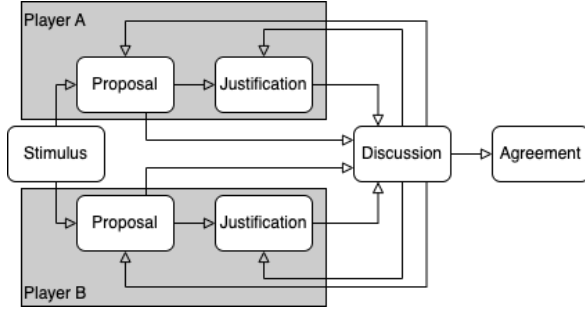
Figure 1: Schema of dialogical classification dialog.

Figure 2: A model of argumentation and explanation in dialog proposed by Bex and Walton (2016, Fig. 4)

likely to help humans make better decisions than they would individually.

Even though some data exists from which such agents can be modeled, it consists largely of only textual data that deal with different topics of discussion, i.e. there is no visual dialog context. In our effort to build explanatory agents that can deal with multimodal input we introduce two new tasks: WORDLE with images, based on the popular Wordle word game, and a collaborative image classification task that comes in two variants: human-human (which we call JUJU for "judgment justification") and human-agent (which we call SIJU for "single-player justification"). We also present data from a naturally occurring dialogical classification setting in discussion forums (below called FORUM). The games we propose differ in their settings and can be used to test the ability to make associations between words and images, and the ability to give and evaluate reasons for proposals.

In the next section, we go into detail about how we intend to extend Miller (2019)'s and Walton (2009)'s models in the context of dialogical classification. We then describe the data collection settings (Section 3) that we use to elicit reasoning chains in interactions. We present example dialogs along with qualitative analyses in Sections 3.3 and 3.4. We end by reviewing related work (Section 5) and summarizing our conclusions with an outlook on future work (Section 6).

2 Collaborative Explanation

Rather than looking at explanations in a general sense, we want to constrain ourselves to explanations of a specific type: Explanations for a classification decision, i.e. given a label $l \in L$ and an input instance $i \in I$, where L is a set of labels and I is a set of images, we want to see dialog around the question "Why was l assigned to i?".
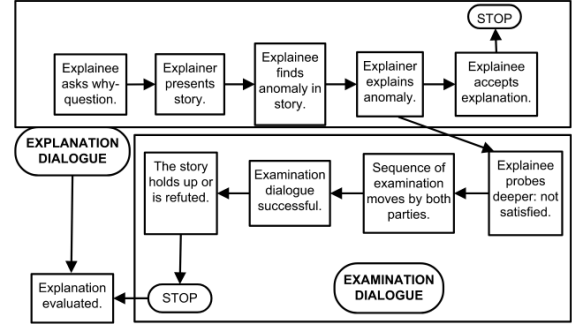
More specifically, we wish to see dialog that

fulfills the following constraints:

- *Grounded justifications*: Justifications for a classification label must be grounded in the input data, i.e. when presented with a label, the explanation must refer to (features of) the particular instance and the domain of the classification.

- *Reasoning towards an agreement*: Rather than explaining a past decision, e.g., "Why did you do this?", we want to see how a decision agreement about a future event is reached, e.g., "What decision should we make and why?". These reasoning chains typically start with a proposal for a solution that is then explained for an explainee but the explainee can make a counter-proposal that can become the solution.

- *Symmetric roles*: We are looking for scenarios in which neither dialog participant knows the ground truth. We are particularly interested in seeing how humans reach a conclusion based on the input that they see and what it takes for one player to agree to the other and when counter-proposals are made.

Looking at Bex and Walton (2016)'s model of explanation dialog that we show in Figure 2, we can see how explanations are initiated by a why-question and how an examination subdialog happens when the explainee requires more insight into the explainer's reasoning. In this model, the dialog ends when either the explainee accepts the explainer's justifications or the explanations (the *story*) are refuted. Table 2 shows some examples of data that Madumal et al. (2019) have gathered to empirically verify the model. From these examples we can see that the topics of discussion in this data is rather abstract and cannot be grounded in the

Task/Data	FORUM	SIJU	JUJU	WORDLE
Sources/Collection method	Reddit, Whatbird	slurk, AMT, Prolific	slurk, AMT	slurk, Prolific
Images	Reddit, Whatbird	CUB	CUB	ImageNet, WikiCommons
Class descriptions	—	Whatbird	Whatbird	—
# dialogs	400	38	11	8

Table 1: Summary of collected data and sources. Appendix A lists details about the data sources.

Question	Source
Over time, did you go, "I need to think this through"?	Journalist Interview
To what extent are you concerned that we might see this problem emerging as a significant challenge [. . .]	
What does that suggest to you?	Chatbot transcript
What are your feelings now?	
How did you guys decide who would walk on the moon first.	Reddit
Why were you wearing this BYU shirt?	
What difference would that have made under the terms of the Hobbs Act?	Court transcript
Why isn't that enough?	

Table 2: Examples of questions that initiate an explanation in the data analyzed by Madumal et al. (2019). Answers to these questions cannot be grounded in the external dialog context.

dialog context and that the roles of explainer and explainee are fixed throughout the dialog, e.g., as interviewer and interviewee. The model and data thus gives insight into the structure of explanation in general and especially shows how the explainer must coordinate with the explainee to reach a common understanding. However the model is not fine-grained enough to capture the phenomenon that we wish to model in which the dialog participants both need to reach an agreement.

Figure 1 shows a schematic diagram of the dialog structure we aim to collect data for. The roles between explainer and explainee can change at any time as players propose a class label for the stimulus. Discussion can lead to players to retract their proposal, e.g., when the other player detects an inconsistency, or to decide on which proposal is better. The dialog ends when players reach an agreement.

3 Data collection settings

In this section, we describe four different settings of dialogical classification, and discuss examples. We start with showing examples of dialogical classification "in the wild": forum discussions in which users are looking to classify images of birds (Section 3.1). We discuss how this setting, while being close to our target, is unsuitable to model dia-

Figure 3: A shortened transcript of a forum interaction.

logical classification as an agreement game. The following three sections describe tasks, data, and methods from our own pilot data collections, for which participants were recruited via crowdsourcing platforms. In two of these settings (JUJU, SIJU, Section 3.3), players interact in the same domain of bird classification. In the fourth setting (WOR-DLE, Section 3.4), players need to take into account several pieces of information (including visual) in order to successfully win a game that they play together. Before describing the settings, we briefly describe the data collection method in Section 3.2.

3.1 Forum data

Dialogical classification happens as part of everyday life. For example, in the domain of bird classification, it is common for people wanting to classify a particular bird that they have spotted in their backyard or during a walk. In specialized forums like WhatBird and general forums like Reddit, anyone

can post an image of a bird and ask others to help them classify the bird. One or more users can then propose a label, explain their choice, and discuss with other users to reach agreement.

Tasks and Data We have collected 130 forum threads from the platform Reddit[2] and 270 threads from the platform WhatBird[3] in which a user posted an image of a bird and asks for help with classifying the image. This data has the potential of showing dialogical classification explanations according to the schema in Figure 1 because a dialog can have a variable number of players in the explainer role. The user posting the request for help is a fixed explainee in this setting but can also ask questions and make proposals. Explanations are likely to be grounded in the specific image uploaded with the request.

For the purpose of this work, we only include forum threads that start with a user posting an image and making a clear classification request that is followed by at least one proposal for a label. Before doing any analysis, the data is cleaned of sensitive information such as usernames and urls. Usernames are translated into neutral identifiers.

Example forum data Figure 3 shows an example forum thread. The dialog starts with A requesting a label and explanation for an image and also giving 3 initial proposals. Two users B and C each provide the same label (turns 2 and 3, with C explicitly agreeing with B) but no explanation. A requests an explanation for the label and B explains the decision in turns 5–7. In this particular case, the explanations also use the principle of exclusion in that they refer to features not present in the image. The dialog ends with A accepting the proposal in turn 8.

Dialog structure Even though both forums are specifically targeted at birders, users in these forums may have different motivations for contributing their knowledge. Providing a label for an image is a courtesy and there is no consequence in being incorrect. In fact, for this data, like for most of the data analyzed by Madumal et al. (2019), no ground truth is available to us as for the correctness of the final answer.

In order to get insight into whether this data exhibits the structure of dialogical classification, we

#	Example
1	Great Potoos are distinctly more white/pale-colored than this bird and they have black eyes.
2	Too large to be Cattle Egrets, and I would have discerned the yellow bills.
3	The white throat really is a great field mark for White-throated Sparrows; it's distinctive even if they have the drab/tan head stripes
4	Goldfinches have shorter more conical beaks compared to the Scarlet Tanager's long slightly curved beak.
5	Even at this angle, it has a slight recurve

Table 3: Examples of forum contributions that reject a proposal or explanation.

look for contributions in the threads in which a user rejects a proposed label or explanation and thus starts a detailed discussion about whether a label matches an instance. We find more than 500 such instances in the 400 dialogs and show examples in Table 3. In the examples, explainers go into detail about why they reject a label and seem to be using a similar strategy of describing concept boundaries that Myrendal (2019) has described for discussions around word meanings: The explanations contrast features of different classes with each other in reference to the stimulus image. These contrasts naturally contain many negations, e.g., by mentioning what an instance of another label would look like (cf. Table 3). While this strategy is valid and interesting to analyze in the future, we are looking for explanations that can be grounded more directly in the image or dialog context, i.e. for positive evidence rather than negative evidence. In the examples, explainers very often reference prototype birds that are not immediately available in the data. As such, these forum interactions are an interesting next step in which a classification model must also learn to reason about the class representation it has built. As a first step however, we work towards the settings where the explanations can be grounded in the immediate context. In addition, this forum data cannot be extended with additional threads on demand as can be done with the tasks we present in the following sections.

3.2   Collecting chat data

For the following three tasks (SIJU, JUJU, and WORDLE), we have used the slurk chat framework (Schlangen et al., 2018; Götze et al., 2022) to set up a chat environment in which players play the game via their internet browser. In each task, players see a chat interface, the game instructions and any game-specific visual material, such as images.

Examples of the interfaces are shown in Figure 4.

An automated game bot that we call *GameMaster* helps the players navigate and informs them about invalid actions and the game score. In the case of the SIJU task, the *GameMaster* is also the player's dialog partner, asking questions to elicit more explanations.

We collect interactions via the platforms Amazon Mechanical Turk and Prolific. Table 1 shows our sources for materials and participants. All data was collected in English. All chat logs are stored on a local server at the authors' institution. Worker IDs are connected to chat ids via a token that players obtain after they finish playing. This allows us to track players who play the game repeatedly and potential repetitions of the same images or words.

We pay workers an average of about $13.00 per hour over all collected data.

3.3 SIJU and JUJU: Judgment Justification

As we have described in Section 3.1, the naturally existing forum data contains aspects of the dialogical classification structure that we aim at but still contains many explanations that cannot be grounded in the immediate dialog context. In this section, we describe two variants of a game in which we constrain the dialog context in a way that allows players to use contrastive explanations that can still be grounded in the context.

We create two variants of a game in which players are tasked to match an image with a description: In the setting that we call JUJU, two players are tasked to create a mapping between 3 images and 3 descriptions. In the setting that we call SIJU, we replace one of the dialog participants with a bot agent that takes the role of a critic. Collecting interaction data for synchronous settings poses an additional challenge in timing participants. With the SIJU setting, we want to investigate whether the illusion of a dialog can be created to the extent that a player will elaborate on their decision further.

JUJU Task Two players are shown 3 images of birds and 3 class descriptions and are tasked to create a mapping between the images and the descriptions. Figure 4 (top right) shows the visual dialog context that participants see. Both images and descriptions are labeled for easy reference (A/B/C and (1/2/3, respectively). In order to avoid the players making decisions by exclusion, the mapping need not include all the images or all the descriptions, there can be images without descriptions and

descriptions without images and the players are informed about this. One round ends when the players both enter their joint decision in free text. The game ends after three rounds.

The *GameMaster*'s role is to keep track of time and contributions and keep the players informed about the state of the game. For example, the *GameMaster* will not accept decisions without prior discussion between the players.

SIJU Task In this variant, the player's task is to decide whether a description fits an image. The player only sees 1 image and 1 description, Figure 4 (top) shows the interface. The game first asks "Does the description fit the image?" and the player answers by clicking one of the buttons yes, no or maybe.

The *GameMaster* that has both the roles of game manager and dialog partner then asks for a justification of the decision. Players click next when they think they have explained enough. The game ends after three pairs.

The *GameMaster*'s role is to challenge the player in their justification by using simple checks of the player's input. The *GameMaster* asks for additional explanation when the explanation falls below a minimum length of 20 characters, fewer than 10% of the features mentioned in the description were taken up in the explanations, or when the explanation is a substring of the description, i.e. the player used copy/paste to answer.

For the images and descriptions, we use the USC-Birds Dataset (Welinder et al., 2010). Besides the bird images, the dataset contains symbolic attribute representations for each bird species. This information is used by the bot to determine which attributes have been mentioned and also allows us to create image-description pairs that have a substantial overlap in order to create pairs where the decision is sufficiently difficult to make.

Collaborative classification in SIJU and JUJU Figures 5 and 6 show examples of collaborative classification dialog in the SIJU dialogs. In Figure 5, the *GameMaster* is not satisfied with the initial explanation and prompts player A to explain further some of the attributes that have not been mentioned before. Player A goes into detail about how the image and description match – making turn #4 a good example of a grounded explanation. In Figure 6, the player goes into detail about the attributes that are mentioned, using positive and

Does the description match the image? Why (not)?

You will see an image of a bird along with a description of a bird species. Decide if the description matches the image by selecting "Yes", "No" or "Maybe".
Please provide as much detail as possible in the chat on the left to explain your reasoning.
When you're finished, click "Next" to move on to the next item.

This medium-sized gull has gray upperparts and white underparts. It has a white head and yellow bill with a black ring near the tip. The wings are gray above, tipped black with white spots, and white below; yellow legs and feet.

Does the description match the image?

| Yes | No | Maybe |

Find out which bird belongs to which species.

You are presented images of birds and descriptions of bird species. Your task is to decide -- together -- which of the birds in the images (if any at all) belong to which of the species described.

(1) Discuss what you are seeing and if any of the birds in the images belong to any of the species described below -- and, if yes, which. For each species there will be at most one matching image.
(2) Once you agree which birds in the images belong to which of the species, you BOTH need to type /decision (don't forget the slash), followed by a brief summary of your decision to indicate that you have agreed on a decision and to finish the game.
(3) Agreeing on a solution is sufficient to finish the game. However, you will receive a bonus only if you get it right!

1) 2) 3)

A) Large, heavy-billed flycatcher with dark olive-brown upperparts, streaked olive-brown sides, and white underparts. Head has slight crest and faint white eye-ring. Wings are dark with two pale bars. Dark tail is relatively short, broad, and slightly notched. Black legs, feet.

B) Small, stocky swallow, dark blue-gray upperparts, pale orange-brown rump, buff underparts. Forehead is white or buff while throat, sides of face are orange-brown. Crown is blue-black, bill is short and black. Tail is dark and squared. Legs and feet are gray.

C) Small flycatcher with olive-gray upperparts, gray breast, and pale yellow belly. Eye-ring is white. The bill has pale lower mandible with dark tip. Legs and feet are black.

Together with your partner, guess the secret 5-letter word. The image gives you a hint, but the connection may be indirect.
In this game mode, only one of you sees the image and needs to describe it to the other person.
You must submit the same guess as your partner, so coordinate! If the guess is not correct, you get more hints.
The earlier you guess the word, the more points you get, so put your heads together!

Green: the letter is in the right position.
Yellow: the letter is in the wrong position.
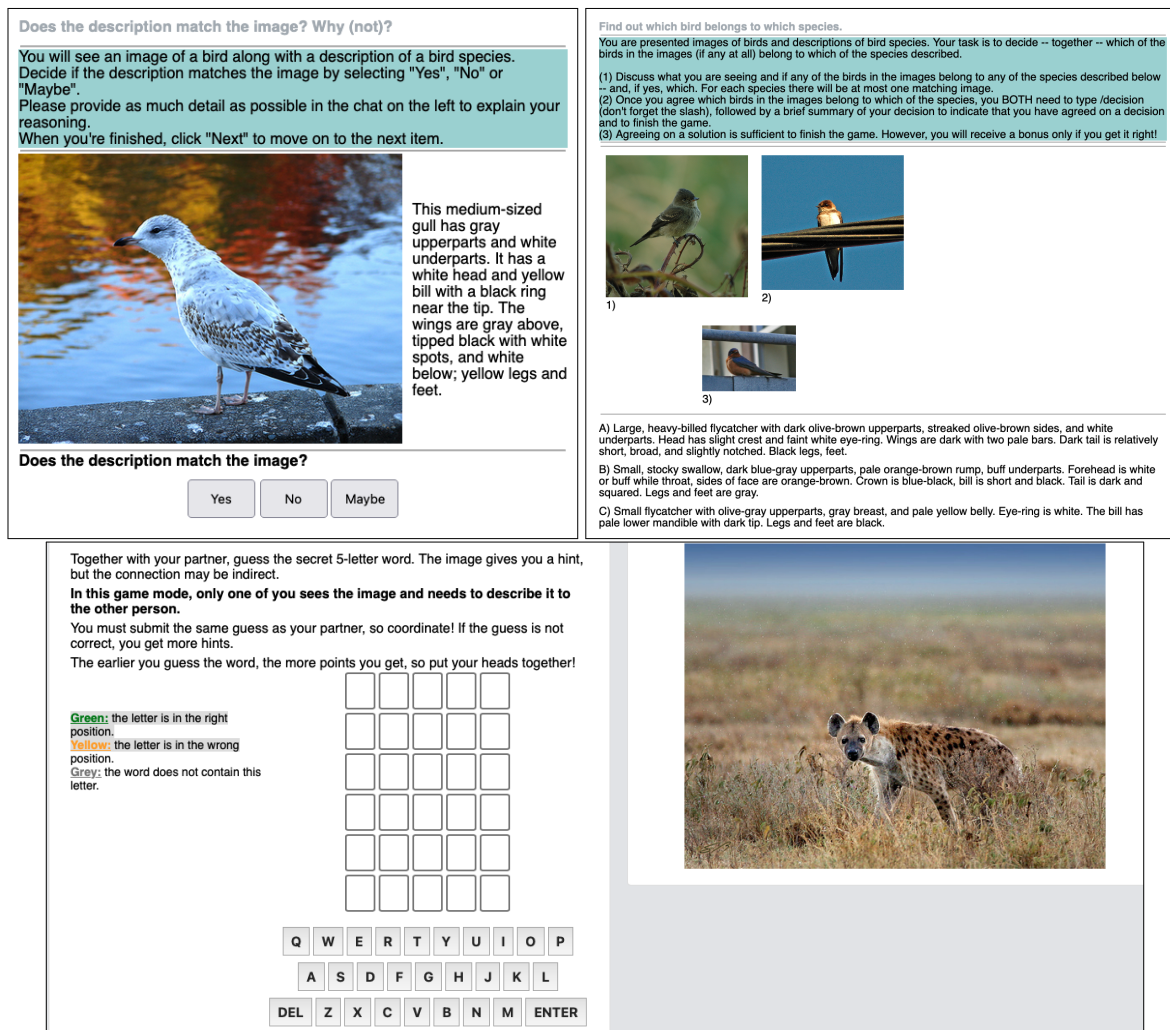Grey: the word does not contain this letter.

Figure 4: The visual interfaces for the SIJU (top left), JUJU (top right), and WORDLE (bottom) games. The interface also includes a chat area that is not shown here. The complete interfaces are shown in Appendix C.

negative references (*"The upperparts do not look reddish."*).

Figure 7 shows an extract of a JUJU dialog in which the players first propose a solution (turns #1 and #2: B: *"3c for sure"*—A: *"yeah I was thinking C could fit 2 or 3"*) and then take turns to explain it in more detail (turns #3 to #6). Here, too, do the players use specific details from the descriptions and ground them in the images, using both the existence and absence of features they see in their explanations (*"B says pale yellow belly and 2 and 3 have white bellies"*, *"1's belly is not pale yellow and it doesn't have an eye ring"*, *"The wings are dark with two white bars"*).

3.4 Wordle with Images

This 2-player game is an extension of the popular Wordle word game[4] in which participants need to

---
[4] https://www.nytimes.com/games/wordle

find a 5-letter word. In addition to cues that specify whether single letters are in the correct position or not, our version also includes an image as a cue for the target word. In this setting, the decision that players make is not a label for the image, but a sequence of letters that needs to adhere to a number of constraints: The target word needs to be related to the image and it needs to fit the letter cues that the *GameMaster* provides as the game progresses. Participants need to agree on their next guess before entering it. The maximum number of guesses is six. The *GameMaster*'s role is to provide letter-based cues and guide players through the mechanics of the game, e.g., informing them when their partner has entered a guess. Figure 4 (bottom) shows a screenshot of the visual game interface.

For our pilot data collection, we have manually mapped images with target words. In order to elicit meaningful dialog, the mapping must be

Medium, active warbler with black upperparts and hood, distinctive orange-red patches on wings, sides, and long, fanned tail, and white underparts. Bill, legs, and feet are black.

#	A	B
1	[A⟩B] The description matches the image – all components have a corresponding part to the image. You could disagree on the disambiguation of Warbler but the colour, pattern, plummage and characteristics of the bird match the description.	
2	[A⟩B] ACTION: BUTTON-NEXT	
3		[B⟩A] I'm not convinced yet. What do you think about the other attributes, such as the side and the feet?
4	[A⟩B] The description states that the sides have distinctive orange-red patches. The image shows that the sides have distinctive orange-red patches. The description states that the feet are black. The image shows that the feet are black.	

Figure 5: An extract of a SIJU interaction. This example shows how the bot prompts the player to be more precise in their explanation. The player then adds more details that can be grounded in the image details.

sufficiently indirect (showing an image of a piano for the target word *piano* does not require the players to discuss more than their agreement). We also add difficulty by showing the image to only one of the players so that this player needs to describe details of the image to their partner.

Collaborative classification in WORDLE   In the WORDLE task, the players now need to observe additional constraints: The target word must be related to the image in some way, it must have exactly 5 letters, and it must adhere to the letter-based feedback. Figure 8 shows an extract in which the players discuss one aspect of the constraints – the letter-based feedback. This subdialog is a good example of one player (A) verbalizing the visual feedback by making a proposal for what the next guess must look like (*"None of the other letters are in the word. Just C."*). The other player B then takes up the proposal by rejecting it because it has misinterpreted the feedback (*"there is an a just in a different position"*). A accepts the counter-proposal in turn #5 and makes a new, more specific proposal in turn #6.

Medium-sized wren with rufous upperparts and buff underparts. Eyebrows are white, wings and tail are dark barred with white flecks. Throat and chin are white. Bill is decurved. Legs and feet are pink-gray.

#	A	B
1	[A⟩B] Medium sized is plausible as this looks larger than the average wren. White-ish eyebrows are visible and throat and chin are arguably white. Bill is decurbed.	
2	[A⟩B] I don't know what rufous means, but I'm guessing reddish. The upperparts do not look reddish. Buff underparts are evidence. Dark wings are evident, but white flecks are not. Tail, legs and feet are not visible.	

Figure 6: An extract of a SIJU interaction.

4   Collaborative explanation models

In the previous section, we have shown data examples of different classification settings that all show aspects of the dialogical classification explanations we aim to capture. Specifically, we see how the dialog partners propose solutions and reason about details in images in descriptions in the SIJU and JUJU settings, using both the presence and absence of feature values in an image (*grounded justifications*). In the JUJU and WORDLE settings, we can see how players make proposals, explain them, and agree on them towards the game goal (*reasoning towards agreement*). Especially in the WORDLE setting, the larger number of constraints on the solutions seems sometimes to elicit incorrect proposals that the other player can counter (cf. the example in Figure 8). Also in the JUJU and WORDLE settings, players' roles are equal, meaning that they alternate between being explainer and explainee since neither has the game solution (*symmetric roles*).

For classification models to give dialogical explanations, what the data in the SIJU setting shows can be considered a minimum capability: On request, a model must be able to go into more detail with an initial explanation, mentioning additional features and possibly admitting that certain features cannot be determined from a given image. The modeling efforts of Li et al. (2018) and Park et al. (2018) are close to this capability, however they lack either the continued dialogical explanation or

A	B	C
Large, crested flycatcher with olive-green upperparts. Head, throat, and upper breast are gray, belly is yellow, and undertail coverts are lemon-yellow. Bill is heavy and black. Wings are dark with rufous patches. Tail is rufous.	Small flycatcher with olive-gray upperparts, gray breast, and pale yellow belly. Eye-ring is white. The bill has pale lower mandible with dark tip. Legs and feet are black.	Medium-sized flycatcher with dull olive-gray upperparts and pale olive-gray underparts. Head has darker cap and slight crest. The wings are dark with two white bars.

#	A	Game Master	B
1		[B⟩A] 3c for sure	
2	[A⟩B] yeah I was thinking C could fit 2 or 3		
3	[A⟩B] A and B don't seem to fit any of them		
4	[A⟩B] B says pale yellow belly and 2 and 3 have white bellies. 1's belly is not pale yellow and it doesn't have an eye ring so I don't think B fits any of them		
5		[B⟩A] slight crest. The wings are dark with two white bars. was what has me thinking 3c	
6		[B⟩A] 1 and 2 have no crests on their heads that I see	
7	[A⟩B] I agree		
8	[A⟩B] so for sure 3c		
9		[B⟩A] yeah, 1 and 2 are confusing	

Figure 7: An extract of a JUJU interaction.

#	A	Game Master	B
	[GM⟩A,B] SUBMIT GUESS: cakes – FEEDBACK: C A K E S		
1	[A⟩B] OK. It's not cakes. But it starts with C		
2	[A⟩B] None of the other letters are in the word. Just C.		
3		[B⟩A] there is an a	
4		[B⟩A] just in a different position	
5	[A⟩B] Oh yeah, sorry.		
6	[A⟩B] Try Cotta		

Figure 8: An extract of a WORDLE interaction.

the questions that the model has to answer are more specific than a general "I need more explanation". The JUJU setting adds more advanced capabilities in which an explainee can make an own proposal that a model must be able to evaluate against its own beliefs. Finally, the WORDLE setting creates a solution space that is constrained by game rules (5 letters, a limited amount of steps) and dynamically changing context (letter-based feedback), as well as a visual input. This latter setting requires advanced reasoning skills and strategic game play in order to stay within the maximum number of allowed steps.

5 Related Work

A growing body of research investigates natural language explanations in the context of classification decisions. Wiegreffe and Marasovic (2021) have compiled an overview of datasets, 10 of which include free-text explanations in a variety of classification tasks that involve visual as well as textual input. Two of these datasets are extensions to the Visual QA task (Antol et al., 2015) in which a model must answer consecutive questions about an image (Li et al., 2018; Park et al., 2018), giving specific elaborations for an answer. Others include explanations for particular action decisions in a given context, e.g., for self-driving cars (Kim et al., 2018) or in a controlled game setting (Ehsan et al., 2019). However none of these datasets include multi-turn fine-grained negotiations and argumentation, or allow the original model decision to be changed during interaction.

There do exist datasets and analyses that put the focus on the process of collaboratively reaching a conclusion or agreement, rather than explaining a specific decision or output. For example, Myrendal (2019) details the collaborative process of negotiating word meanings using the example of online forums. In these interactions, the participants are not necessarily reaching agreement, but give detailed explanations for the aspects of a word meaning that they find relevant in a particular situation. The FORUM data we have shown in Section 3.1 is similar in this respect. Similarly, Madumal et al. (2019) have analyzed dialog data from different explanation settings, including settings in which either the

explainer or the explainee is an automated agent, empirically affirming Walton (2009)'s theoretical formalization of dialog structure.

Madumal et al. (2019)'s work makes visible that explanatory interactions contain a component of argumentation, in which claims can be questioned and counter-arguments be proposed. Computational argumentation has its own line of research that incorporates aspects that are important for modelling an explainer, e.g., Wachsmuth and Alshomary (2022) build a model that can generate arguments that are user-dependent, taking into account the explainee's previous beliefs.

While these settings assume specialized roles for the interaction participants, these roles could change throughout a conversation. In the domains of negotiation and collaborative problem-solving, participants are both explaining and being explained to. Both types of dialog require participants to justify their proposals or make counter-proposals to achieve either a common or individual goal. Data exists based on in-lab data collections (Petukhova et al., 2016; Karadzhov et al., 2022a) and Wikipedia data (De Kock and Vlachos, 2021). DeliData (Karadzhov et al., 2022a) is very similar to the WORDLE game we are proposing here in that participants are explaining fine-grained options for a solution, but there is no image context.

6 Conclusion and Outlook

All three settings that we have shown elicit aspects of dialogical classification explanation. In the next step, we plan to collect a larger number of dialogs from each game and annotate them systematically in order to add a quantitative analysis to the qualitative examples we have shown here.

Acknowledgments

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Floris Bex and Douglas Walton. 2016. Combining explanation and argumentation in dialogue. *Argument & Computation*, 7(1):55–68.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA. Association for Computing Machinery.

Christine De Kock and Andreas Vlachos. 2021. I Beg to Differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online. Association for Computational Linguistics.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. 2019. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 263–274.

Jana Götze, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann, and David Schlangen. 2022. The slurk interaction server framework: Better data for better dialog models. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4069–4078, Marseille, France. European Language Resources Association.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022a. DeliData: A dataset for deliberation in multi-party problem solving. *arXiv:2108.05271*.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022b. What makes you change your mind? An empirical investigation in online group decision-making conversations. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 552–563, Edinburgh, UK. Association for Computational Linguistics.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 577–593, Cham. Springer International Publishing.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 570–586, Cham. Springer International Publishing.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS

'19, pages 1033–1041, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38.

Jenny Myrendal. 2019. Negotiating meanings online: Disagreements about word meaning in discussion forum communication. *Discourse Studies*, 21(3):317–339.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Volha Petukhova, Christopher Stevens, Harmen de Weerd, Niels Taatgen, Fokie Cnossen, and Andrei Malchanau. 2016. Modelling Multi-issue Bargaining Dialogues: Data Collection, Annotation Design and Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3133–3140, Portorož, Slovenia. European Language Resources Association (ELRA).

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

David Schlangen, Tim Diekmann, Nikolai Ilinykh, and Sina Zarrieß. 2018. Slurk – A Lightweight Interaction Server For Dialogue Experiments and Data Collection. In *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial / Semdial 2018)*, Aix-en-Provence, France.

Henning Wachsmuth and Milad Alshomary. 2022. "Mama Always Had a Way of Explaining Things So I Could Understand": A Dialogue Corpus for Learning to Construct Explanations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 344–354, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Douglas Walton. 2009. Objections, Rebuttals and Refutations. *OSSA Conference Archive*.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-201, Caltech.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

A   Data usage

We have used material from a number of sources. For the SIJU and JUJU tasks, we have used texts and images from `https://forums.whatbird.com` with their explicit permission that was granted via email. Images for these tasks also come from the CUB dataset (Welinder et al., 2010). For the WORDLE tasks, we have used images from ImageNet ILSVRC2012 dataset (Russakovsky et al., 2015) and Wikimedia Commons (`https://commons.wikimedia.org/wiki/Main_Page`). The corresponding target words have been selected manually for now.

B   Example dialogs

Figures 9, 10, and 11 show example dialogs for the SIJU, JUJU, and WORDLE games, respectively.

C   Chat interfaces

Figure 12 shows the complete user interfaces for the three data collections SIJU, JUJU, and WORDLE at a larger scale.

#	A	Game Master	B
	[Task intro]		
		[B⟩A] Great! Let's begin.	
1	[A⟩B] ACTION: BUTTON-AGREE		
2		[B⟩A] Really? I don't agree. Which attributes in the text make you think so? Convince me!	
3	[A⟩B] The description matches the image - all components have a corresponding part to the image. You could disagree on the disambiguation of Warbler but the colour, pattern, plummage and characteristics of the bird match the description.		
4	[A⟩B] ACTION: BUTTON-NEXT		
5		[B⟩A] I'm not convinced yet. What do you think about the other attributes, such as the side and the feet?	
6	[A⟩B] The description states that the sides have distinctive orange-red patches. The image shows that the sides have distinctive orange-red patches. The description states that the feet are black. The image shows that the feet are black.		
7	[A⟩B] ACTION: BUTTON-NEXT		
8		[B⟩A] Okay, let's get you to the next round. 2 to go!	
9	[A⟩B] ACTION: BUTTON-AGREE		
10		[B⟩A] Really? I'm not so sure. Which attributes in the text make you think so? Convince me!	
11	[A⟩B] The description matches the image - all components have a corresponding part to the image. The colour, pattern, plummage and characteristics of the bird match the description.		
12	[A⟩B] ACTION: BUTTON-NEXT		
13		[B⟩A] I'm still not sure. What about the other attributes, like the head and the eyes?	
14	[A⟩B] You could disagree that the head has a purple sheen, as some individuals have a higher sensitivity to colour. However, the eyes are very dark. Unmistakeably so.		
15	[A⟩B] ACTION: BUTTON-NEXT		
16		[B⟩A] Okay, let's get you to the next round. 1 to go!	
17	[A⟩B] ACTION: BUTTON-MAYBE		
18		[B⟩A] Alright. Are there any features that you ARE certain about? Which attributes in the text make you unsure?	
19	[A⟩B] Small. Yellow Underparts. chestnut-brown ear patch, black crown.		
20	[A⟩B] ACTION: BUTTON-DONE		
21		[B⟩A] That was the last round. Thank you for participating!	

Image			
Description	Medium, active warbler with black upperparts and hood, distinctive orange-red patches on wings, sides, and long, fanned tail, and white underparts. Bill, legs, and feet are black.	Small blackbird with purple sheen on head, back, breast. Eyes are dark. Black bill, legs, feet.	Small warbler, olive-yellow upperparts, thick, black streaks on yellow underparts. Bright yellow face, chestnut-brown ear patch, black crown. Wings are dark with large white patches.

Figure 9: A transcript from a SIJU interaction. Text and description match for the first two images. The mismatch in the third pair is a difficult instance as parts of the description are correct.

#	A	Game Master	B
	[Task intro]		
	[GM)A,B] Yay! The game will begin now.		
1	[A)B] This is a test dialog.		
2		[B)A] None of these is a gul, so A) doesn't go with any picture	
3	[A)B] B) neither		
4	[A)B] none of these match		
5		[B)A] DECISION: NO MATCH	
6		[GM)B] Are you sure? Please discuss some more!	
7	[A)B] DECISION: NO MATCH HERE		
8	[GM)A] Are you sure? Please discuss some more!		
9	[A)B] c could go with 3, but 3 doesn't have a black face		
10		[B)A] or a pink bill	
11		[B)A] it's more grayish	
12	[A)B] DECISION: NO MATCH		
13	[GM)A] Let's wait for your partner to also type /decision.		
14		[GM)B] Your partner thinks that the two of you have made a decision. Type /decision and a brief explanation if you agree.	
15		[B)A] DECISION: NO MATCH	
16	[GM)A,B] Ok, let's get the two of you the next level. 2 to go!		
17	[A)B] d goes with 4		
18		[B)A] the others dont seem to match	
19	[A)B] none of these have a white neck		
20		[B)A] and there's no description that's saying anything about a bird with a red head	
21	[A)B] nothing with bright yellow either		
22	[A)B] DECISION: 4D		
23	[GM)A] Are you sure? Please discuss some more!		
24	[A)B] yes I'm sure		
25	[A)B] DECISION: 4D		
26	[GM)A] Are you sure? Please discuss some more!		
27	[A)B] yes sure		
28		[B)A] I'm sure too	
29		[B)A] DECISION: 4D	
30		[GM)B] Let's wait for your partner to also type /decision.	
31	[GM)A] Your partner thinks that the two of you have made a decision. Type /decision and a brief explanation if you agree.		
32	[A)B] DECISION: 4D		
33	[GM)A,B] Ok, let's get the two of you the next level. 1 to go!		
34		[B)A] ooh tricky	
35		[B)A] c3?	
36	[A)B] yes agree		
37	[A)B] and 1a		
38		[B)A] test	
39		[B)A] DECISION: 1A 3C	
40		[GM)B] Are you sure? Please discuss some more!	
41	[A)B] yes sure		
42	[A)B] sure		
43	[A)B] DECISION: 1A 3C		
44	[GM)A] Let's wait for your partner to also type /decision.		
45		[GM)B] Your partner thinks that the two of you have made a decision. Type /decision and a brief explanation if you agree.	
46		[B)A] DECISION: 1A 3C	
47	[GM)A,B] The game is over! Thank you for participating!		

Figure 10: A transcript from a JUJU interaction.

| #	A	Game Master	B
1	[GM⟩A,B] Welcome to Wordle with Images		
2	[GM⟩A,B] Let's start with the first of 1 images		
3	[A⟩B] Hello!		
4		[B⟩A] hello	
5		[B⟩A] my guess is the word music	
6	[A⟩GM] GUESS: music		
7	[GM⟩A] Let's wait for your partner to also enter a guess.		
8		[GM⟩B] Your partner thinks that you have found the right word. Enter your guess.	
9		[B⟩GM] GUESS: music	
10	[GM⟩A,B] SUBMIT GUESS: music – FEEDBACK: M **U** S I C		
11	[A⟩B] Do you have an image? I'm just seeing normal wordle		
12		[B⟩A] yes i have an image, its of a homeless looking guy playing a flute with a dog	
13	[A⟩B] Audio?		
14	[A⟩B] Wait no I		
15		[B⟩A] no aduio	
16	[A⟩B] I've got no clue		
17	[A⟩B] Any other details in the image?		
18		[B⟩A] they arestanding on a crumpled blanket on a cobblestone like sidewalk the dog is a dalmation	
19		[B⟩A] the guy has a sleevless tank top and shaggy hair	
20	[A⟩B] Maybe buddy? Because of his dog?		
21		[B⟩A] yes probably buddy	
22		[B⟩GM] GUESS: buddy	
23		[GM⟩B] Let's wait for your partner to also enter a guess.	
24	[GM⟩A] Your partner thinks that you have found the right word. Enter your guess.		
25	[A⟩GM] GUESS: buddy		
26	[GM⟩A,B] SUBMIT GUESS: buddy – FEEDBACK: B **U** D D **Y**		
27	[A⟩B] Okay so we've got U and Y		
28	[A⟩B] B and D are wrong		
29		[B⟩A] i typed in funky before i switched to buddy	
30	[A⟩B] Oooo funky could work, because of the music		
31	[A⟩GM] GUESS: funky		
32	[GM⟩A] Let's wait for your partner to also enter a guess.		
33		[GM⟩B] Your partner thinks that you have found the right word. Enter your guess.	
34		[B⟩GM] GUESS: funky	
35	[GM⟩A,B] SUBMIT GUESS: funky – FEEDBACK: F **U** N K **Y**		
36	[A⟩B] Let's guess Funky		
37	[A⟩B] Lmao I'm a bit confused now		
38	[A⟩B] Perhaps the word is unrelated to the picture?		
39		[B⟩A] maybe	
40	[A⟩B] Puppy? Yuggy?		
41	[A⟩B] Ruggy*		
42	[A⟩B] Not even sure Ruggy is a word		
43		[B⟩A] puppy or pushy?	
44	[A⟩B] Can't be S, used that in Music		
45	[A⟩B] Wanna try Puppy?		
46		[B⟩A] sure	
47	[A⟩GM] GUESS: puppy		
48	[GM⟩A] Let's wait for your partner to also enter a guess.		
49		[GM⟩B] Your partner thinks that you have found the right word. Enter your guess.	
50		[B⟩GM] GUESS: puppy	
51	[GM⟩A,B] SUBMIT GUESS: puppy – FEEDBACK: **P** **U** **P** **P** **Y**		
52	[GM⟩A,B] YOU WON! For this round you get 10 points. Your total score is: 10		
53	[GM⟩A,B] The game is over! Thank you for participating!		
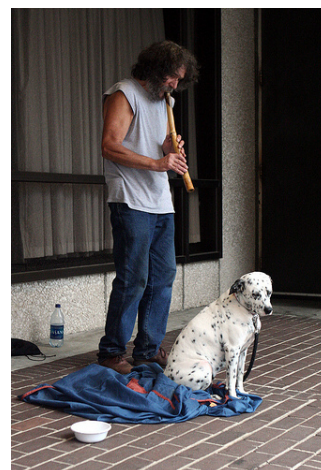
Figure 11: A shortened transcript from a Wordle interaction. The image is taken from the ImageNet ILSVRC2012 dataset (Russakovsky et al., 2015).
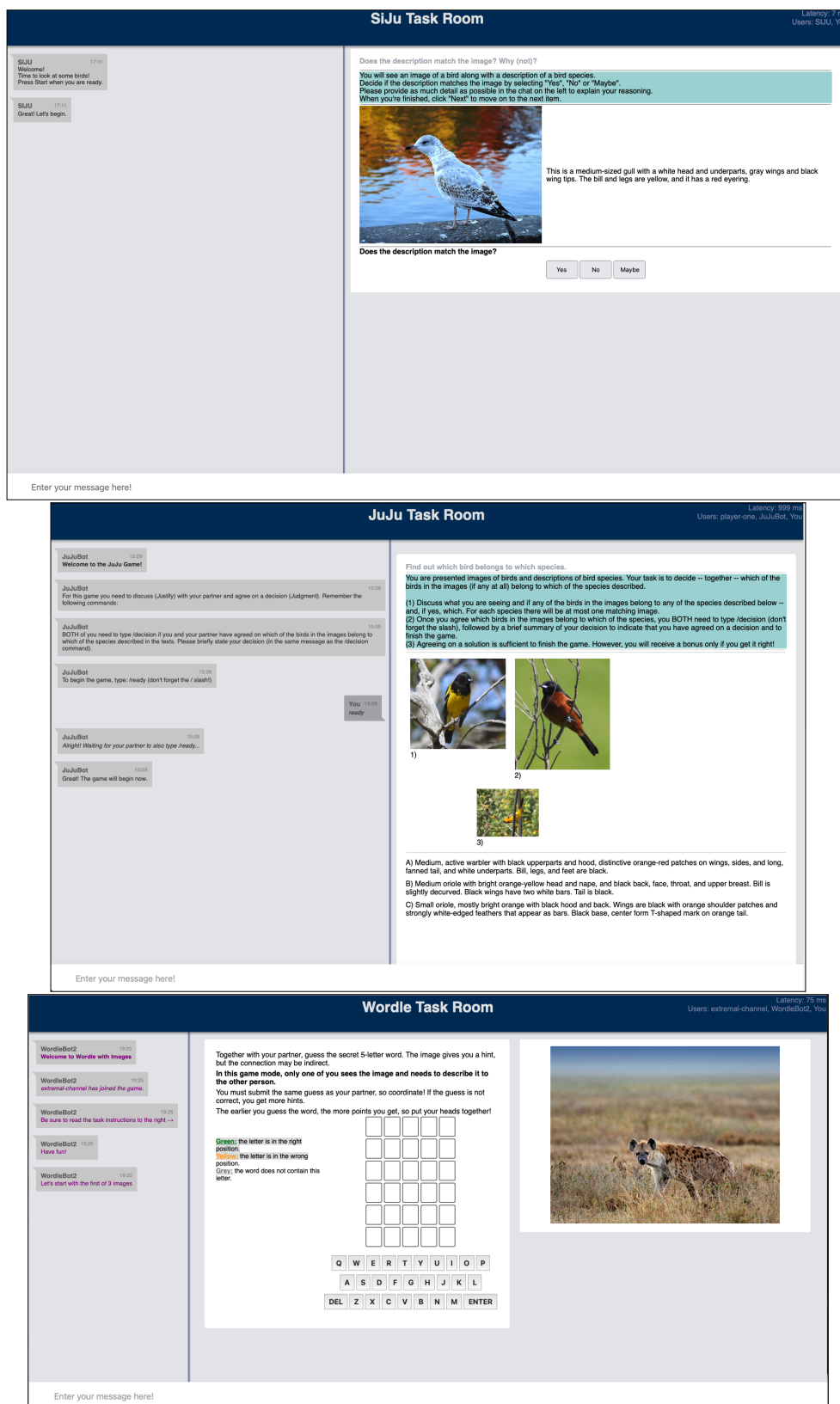
Figure 12: The visual interfaces for the SIJU (top), JUJU (middle), and WORDLE (bottom) games.