PRAGEU: Praise Generation Using End-to-End Neural Models

Anonymous ACL submission

Abstract

Humans employ a range of conversation strategies during conversation to achieve multiple goals in dialogue. One of such strategies is a positive evaluation made by a person of another's attributes, known as *Praise*. State of the art neural dialogue models attempt to engage human users without taking into account this conversation strategy. Hence in this work, we present a method of generating praise using state of the art natural language generation models. We achieve this by collecting a dataset using amazon mechanical turk (AMT) using Persona-Chat and create a new corpus called Praise-on-Persona (POP) and fine-tune various models to generate praise. Our results show that large language models can learn to link an attribute and a praise associated with it, such a a *Professor* and *Intelligence* or *PhD* and *hard work*.

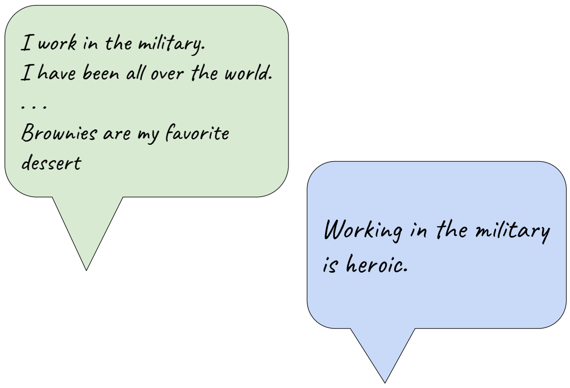

Figure 1: Praise on Persona example from our POP Corpus gathered by using Persona-Chat (Zhang et al., 2018)

1 Introduction

Pursuing more than one objective while having a conversation is a key property of human dialogue (Tracy and Coupland, 1990). A dialogue can address propositional goals to convey information, interactional goals ensure that the conversation proceeds smoothly, and interpersonal goals to work towards increased rapport (Fetzer, 2003; Cassell and Bickmore, 2003). Kanouse et al. (1981) define praise as "positive evaluation made by a person of another's products, performance, or attributes. where the evaluator presumes the validity of the standards on which the evaluation is based". Praise is a conversational strategy which serves to increase interpersonal cohesiveness (Zhao et al., 2016; Delin and Baumeister, 1994),. Praise also functions as speech act which creates or reinforces solidarity between the interlocutors (Wolfson and Manes, 1980). Since praise is commonly given and generally does not require much history or association with the current conversation (Wolfson and Manes, 1980), the praise-generation module, that we create here, can be called-in as a separated module for dialogue in neural end-to-end dialogue dialogue systems. Large language models (LLM) can interact with humans and generate coherent responses, which are very useful in many cases, such as code generation, question answering etc. However, LLM are not good at creating a relationship with the user, by using a conversation strategy such a *Praise*. Towards this end, in this work, we work towards the development of a generation module for praise. Hence, we design a pipeline to utilise pre-trained models to generate praise. Although *Praise* has many affects such as as for *behaviour reinforcement*, *internal-motivation*, we are concerned with the function of praise in dialogue and the resulting effect on the interlocutors and below we survey praise-theory followed by praise-generation.

In the rest of the paper, we first survey relevant work, followed by description of how we gathered the POP corpus, we then given experimental background and discuss results obtained on automatic and human evaluations.

2 Related Work

This section surveys praise-theory followed by existing work on praise or other conversation strategy generation.

2.1 Praise - Theory

There are many function of praise: *behavioural reinforcement*, which is primarily based on Skinner (1963) model of operant conditioning. Praise has been seen as a means of positive reinforcement and many studies have demonstrated the effectiveness of praise in reinforcing behaviour, in teacher-student (Brophy, 1981; Strain et al., 1983) and parent-child based scenarios (Garland et al., 2008) and work performance (Crowell et al., 1988). *motivation enhancement:* Praise serves as a powerful motivator. Praise theory posits that positive feedback can increase an individual's intrinsic motivation (Deci et al., 1999), enhancing their commitment to a task or goal. Self-Esteem Development: Praise contributes to the development of self-esteem. Erikson (1993) suggests that praise and encouragement during childhood play a vital role in the formation of a positive self-identity. Cultural Variation: Praise is subject to cultural variation. Research by (Kitayama et al., 2006) highlights how cultures differ in their expressions of praise and their impact on self-esteem. However, we are interested in praise's role in *Interpersonal Bonding*: Praise fosters stronger interpersonal relationships. Expressing admiration and appreciation can strengthen social bonds and build trust (Algoe et al., 2010; Wolfson and Manes, 1980). This interpersonal bond is also known by the term *Rapport* (Spencer-Oatey, 2008; Ambady and Rosenthal, 1993) and is built by employing various conversation strategies such as self-disclosure, hedging and praise inter-alia. Hence, in this work we create a module to generate praise.

2.2 Praise Generation

As this study focuses on generating praise in the context of conversation strategies generation, in this section, we review past research to generate conversation strategies. A few past studies have attempted to generated praise, for example in a embodied conversational agents (ECAs) used in computer assisted language learning Wik and Hjalmarsson (2009); Davison et al. (2021). (Soni et al., 2021) utilised re-ranking techniques to generate self-disclosure in DialoGPT and (Abulimiti et al., 2023) generated hedges utilising three text generation models. Within the domain of conversation strategies generation, it is important to investigate how a particular strategy can be generated, within the context of generating them using a LLM. Although, self-disclosure and hedging are strategies that are found in the lower probability tokens while generation, *Praise* simply cannot be generated by re-ranking. Since praise is given on attribute, Personas in Persona-Chat corpus (Zhang et al., 2018) fit our use case. Persona-chat is a widely used dialogue corpus in english. Towards this end, we gather praise on persona and curate the praise-on-persona (POP) corpus.

Persona-Praise Pairs	8939
Persona(min)	10
Persona(max)	78
Persona(median)	27
Praise (min)	3
Praise (max)	123
Praise (mean)	10

Table 1: POP corpus statistics

3 Data Collection - The POP Corpus

As described above, the next step was to collect human generated praise on the basis of persona. We employ workers from amazon mechanical turk (AMT) to gather praise. We only selected workers with a master status and who had english as their first language, to ensure data quality. The instruction that we framed were simple, it was to give a praise, given a persona. We also provided examples of good and bad praise, mentioning that the workers should avoid generic praise and that the praise be grounded in persona. Figure 2 in the appendix is a screen capture showing the precise instructions and Figure 3 shows good and bad examples provided to the AMT workers. It was easier to provide praise on some personas while it was difficult to provide for others. These were usually the case, when the user hasn't revealed much information about themselves or positive evaluation of a quality is not possible. However, in the initial stage of data collection we did not filter persona-chat and trusted the state-of-the-art models to learn to generate praise from the dataset, and as we will see later, the models we employed did learn to generate praise. Table 1 contains the details of the corpus. The corpus will be released as a part of this research work.

4 Experiment

Following the collection of the POP corpus, the next task was to employ suitable models and generation techniques. The task is of generating a Praise Pr, given a Persona Per. The dataset obtained in section 3 was randomly split in 81:09:10 ratio as training, validation and test set. We report the human evaluation on 300 persona-praise pair, 100 for each model.

4.1 Text-to-Text Generation Models

Since this task is similar to learning mapping between a pair of texts, we utilise models that are appropriate for Text-to-Text Generation such as T5 (Raffel et al., 2019), BART (Lewis et al., 2019) and BlenderBot (Roller et al., 2020). These models were trained to learn Text-to-Text representations. Here is a brief description of these models and why we use them. To generate praise, we fine-tune the models below on the POP dataset.

T5 (Raffel et al., 2019) T5 is a encoder-decoder model which converts all NLP problems into a text-to-text format. It is trained on different datasets using teacher forcing. T5 is fine-tunable on both supervised and unsupervised fashion. Hence making T5 suitable for text2text generation tasks such as persona2praise generation. We utilise T5-base for our experiment.

BART (Chung et al., 2022) Bart consist of a seq2seq/machine translation architecture with a bidirectional encoder and a left-to-right decoder. BART is trained on two modelling functions 1) Corrupting tokens and 2) reconstruction of a given text sequence. BART is specially effective for our task due to the second modelling function. We utilise BART-large for our experimentation.

BlenderBot (Roller et al., 2020) Blenderbot uses the Seq2Seq Transformer architecture and is trained on a number of datasets such as Empathetic Dialogue (Rashkin et al., 2018), Persona-chat (Zhang et al., 2018), ConvAI2 (Dinan et al., 2020), and other dialogue datasets and is capable of learning particular kinds of generation, such as praise. This along with Blender-bot being a dialogue model make it appropriate for generating praise.

4.2 Evaluation Metrics

We employ Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) as our primary evaluation metric, as we were interested in finding the recall of overlap between human and generated praise. We specifically use ROUGE1, ROUGE2 and ROUGE L. ROUGE is a case-insensitive metric which measures the recall between reference and predictions by the means of words and/or n-grams. ROUGE1 measures the *unigram* while ROUGE2 measures bi-grams, ROUGE-L is then based on Longest Common Sub-sequence between the reference and model predictions and ROUGE-LSum calculate LCS based on `\n` . In addition to ensure quality, we measure Generation Length.

4.3 Human Evaluation

While automated metrics are necessary to ensure the quality of generation, they do not tell us if the generated text is praise and relevant to the persona. Human evaluation is thus necessary to make sure that the generated text is praise and relevant to the persona. Towards this end, we employ 2 human annotators from the same research organisation we work in by sending email about the task description and ask them to annotate praise as relevant praise, given a persona. Their first task is to evaluate if a given sentence is praise and second to evaluate if it is relevant to the persona. The inter-annotator agreement was found to be 0.8, indicating significant agreement.

	T5	BART	BlenderBot
ROUGE 1	49.23	**54.40**	40.20
ROUGE 2	36.16	**45.23**	30.19
ROUGE-L	47.82	**53.89**	38.12
ROUGE-LSum	47.74	**53.98**	39.43
Gen-Len	11.25	10.07	**26.94**
F1Score (hum-eval)	0.85	**0.94**	0.87
Accuracy (hum-eval)	0.74	**0.89**	0.78

Table 2: Automatic and human evaluation scores of the fine-tuned models on the POP dataset.

5 Results

Table 2 shows the overall results and Table 3 shows example of praise generated. On evaluation, we found that BART outperformed the other models in terms of ROUGE scores, this could be because

Persona:	"I graduated college in 2016. I enjoys kayaking in my free time. I teach kindergarten. My class has 26 kids. I teach at a large school."
T5:	"You must be talented!"
BART :	"You must be a hard worker."
Blender-Bot :	"You must be a hard worker. I admire that. Good for you. Keep up the good work."

Table 3: An Example of Praise from each of our fine-tuned models

BART is pre-trained to reconstruct the original text. The generation length was longest from Blender-Bot as the praise generated from BlenderBot is more conversational and usually contains a greeting at the end. The difference in ROUGE scores between BART and T5 was significant yet numerically near. There were characteristics of praise generated by different models. We noticed that BART produced praise which were of the kind "you must be *X*", where X is a praise of an attribute such as "you must be *hardworking*". This tells us that by fine-tuning the model has learnt a mapping between persona attribute and praise, for example *effort* and *hardworking* ;*skill* and *talented*. Praise generated by Blenderbot was more *conversational*, most praise were suffixed by "I wish you good luck" or "more power to you". Hence the praise generated from BlenderBot was long with a mean length of 26.94. Using larger versions of T5 gave us lower ROUGE scores, this could be because our dataset is relatively small and hence a smaller model learns the representation better. The results of the human evaluation are also consistent with the automatic metrics. We calculate accuracy and F-1 score between human annotators labelling a model output as praise and relevant as 1 and 0 otherwise and model output as 1 since model outputs are always praise. BART outperform T5 and BlenderBot in both accuracy and F1-score with T5 and BlenderBot fairing closely. **Overall we find that fine-tuning on persona-praise pairs enables models to generate praise.**

6 Limitations

This work has several limitations. First, we forcing the generation of praise, which sometimes does ignore a negative attribute about the user. We are aware that in real life conversations, such might not be the case. Second, the corpus collected here is a synthetic corpus and we haven't gathered praise from real world conversation. Third, we have not delved deep into the kinds of praise that exist and have simply generated praise in a black-box way. Future work will include generation of different kinds of praise. We also did not compare the performance of the fine-tune models to prompting based models, in a small experiment however, we employed ChatGPT to generate praise on persona and the praise generated were unnatural and not human-like. However, in future works, we would like to compare the results of in-context learning to fine-tuning in generating praise.

7 Conclusion and Future Work

In this concise research experimentation, we were able to formulate a novel task of generating praise given a persona. We were able to show that text-to-text generation models can indeed learn to generate praise. The PRAGEU module can become part of any end-to-end dialogue system where a federated dialogue manager may signal to invoke the PRAGEU module to generate praise. In future works, we would like to evaluate the effect of praise within a conversation.

8 Ethical Statement

We obtained ethical clearance to collect the dataset. AMT workers were paid paid 0.03 $ per hit. We made sure to approve the tasks promptly so that the workers are paid quickly. For human annotations, two annotators were paid 10€ for their effort. We are also publicly releasing the dataset for wider-research usage. In using generation models, we were careful to avoid generation of inappropriate content and since Praise is about positive evaluation of a attribute, it rarely leads to inappropriate content generation. Finally, it is imperative to declare that the praise is being generated by a machine so that the generation of praise cannot be used by spam-bots to deceive users.

References

Alafate Abulimiti, Chloé Clavel, and Justine Cassell. 2023. How about kind of generating hedges using end-to-end neural models? *ArXiv*, abs/2306.14696.

Sara B Algoe, Shelly L Gable, and Natalya C Maisel. 2010. It's the little things: Everyday gratitude as a booster shot for romantic relationships. *Personal relationships*, 17(2):217–233.

Nalini Ambady and Robert Rosenthal. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology*, 64(3):431.

Jere Brophy. 1981. Teacher praise: A functional analysis. *Review of educational research*, 51(1):5–32.

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User modeling and user-adapted interaction*, 13(1):89–132.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Charles R Crowell, D Chris Anderson, Dawn M Abel, and Joseph P Sergio. 1988. Task clarification, performance feedback, and social praise: Procedures for improving the customer service of bank tellers. *Journal of applied behavior analysis*, 21(1):65–71.

Daniel P Davison, Frances M Wijnen, Vicky Charisi, Jan van der Meij, Dennis Reidsma, and Vanessa Evers. 2021. Words of encouragement: How praise delivered by a social robot changes children's mindset for learning. *Journal on multimodal user interfaces*, 15:61–76.

Edward L Deci, Richard Koestner, and Richard M Ryan. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6):627.

Catherine R Delin and Roy F Baumeister. 1994. Praise: More than just social reinforcement. *Journal for the theory of social behaviour*, 24(3):219–241.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Erik H Erikson. 1993. *Childhood and society*. WW Norton & Company.

Anita Fetzer. 2003. 'no thanks': a socio-semiotic approach. *Linguistik online*, 14(2).

Ann F Garland, Kristin M Hawley, Lauren Brookman-Frazee, and Michael S Hurlburt. 2008. Identifying common elements of evidence-based psychosocial treatments for children's disruptive behavior problems. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(5):505–514.

David E Kanouse, Peter Gumpert, and Donnah Canavan-Gumpert. 1981. The semantics of praise. *New directions in attribution research*, 3:97–115.

Shinobu Kitayama, Batja Mesquita, and Mayumi Karasawa. 2006. Cultural affordances and emotional experience: socially engaging and disengaging emotions in japan and the united states. *Journal of personality and social psychology*, 91(5):890.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Burrhus F Skinner. 1963. Operant behavior. *American psychologist*, 18(8):503.

Mayank Soni, Benjamin R. Cowan, and Vincent P. Wade. 2021. Enhancing self-disclosure in open-domain dialogue by candidate re-ranking. In *International Workshop on Spoken Dialogue Systems Technology*.

Helen Spencer-Oatey. 2008. *Culturally Speaking Second Edition: Culture, Communication and Politeness Theory*. Bloomsbury Publishing.

Phillip S Strain, Deborah L Lambert, Mary Margaret Kerr, Vaughan Stagg, and Donna A Lenkner. 1983. Naturalistic assessment of children's compliance to teachers'requests and consequences for compliance. *Journal of Applied Behavior Analysis*, 16(2):243–249.

Karen Tracy and Nikolas Coupland. 1990. *Multiple goals in discourse*. Multilingual Matters.

Preben Wik and Anna Hjalmarsson. 2009. Embodied conversational agents in computer assisted language learning. *Speech communication*, 51(10):1024–1037.

Nessa Wolfson and Joan Manes. 1980. The compliment as a social strategy. *Research on Language & Social Interaction*, 13(3):391–410.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. 2016. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 381–392, Los Angeles. Association for Computational Linguistics.

A Parameters and Hyper-parameters

This section will enlist the different parameters, not in order, that were used to generate praise. We utilised top K = 200 as limiting the number of top-k's seemed to give us the best praise results. We experimented with other sampling techniques but found top-k to be most effective. Batch size utilised during training was 8 and during testing was 32. The learning rate was 4e-5. The number of epochs was 1 for all models for fair comparison, we did not fine-tune beyond 1 epoch to avoid overfitting. Other parameters could be found in the code attached. All the models were utilised from HuggingFace. All the models were fine-tuned in less than 30 minutes with 20 minutes the average on T5 GPUs on google collab.

B Data Collection - Amazon Mechanical Turk Screenshots

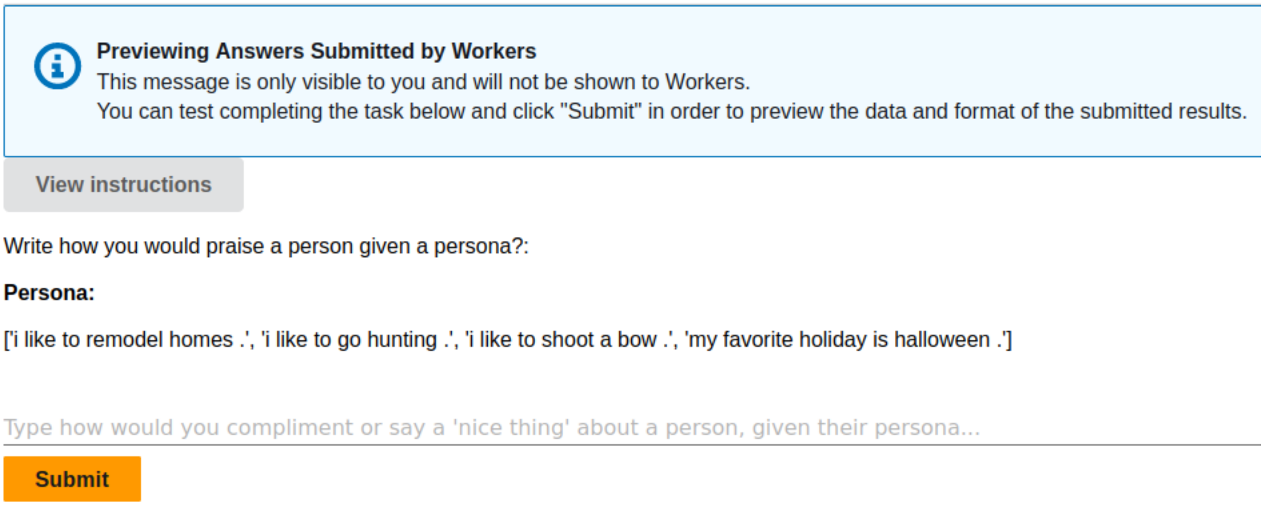

Figure 2: User Interface containing Instructions

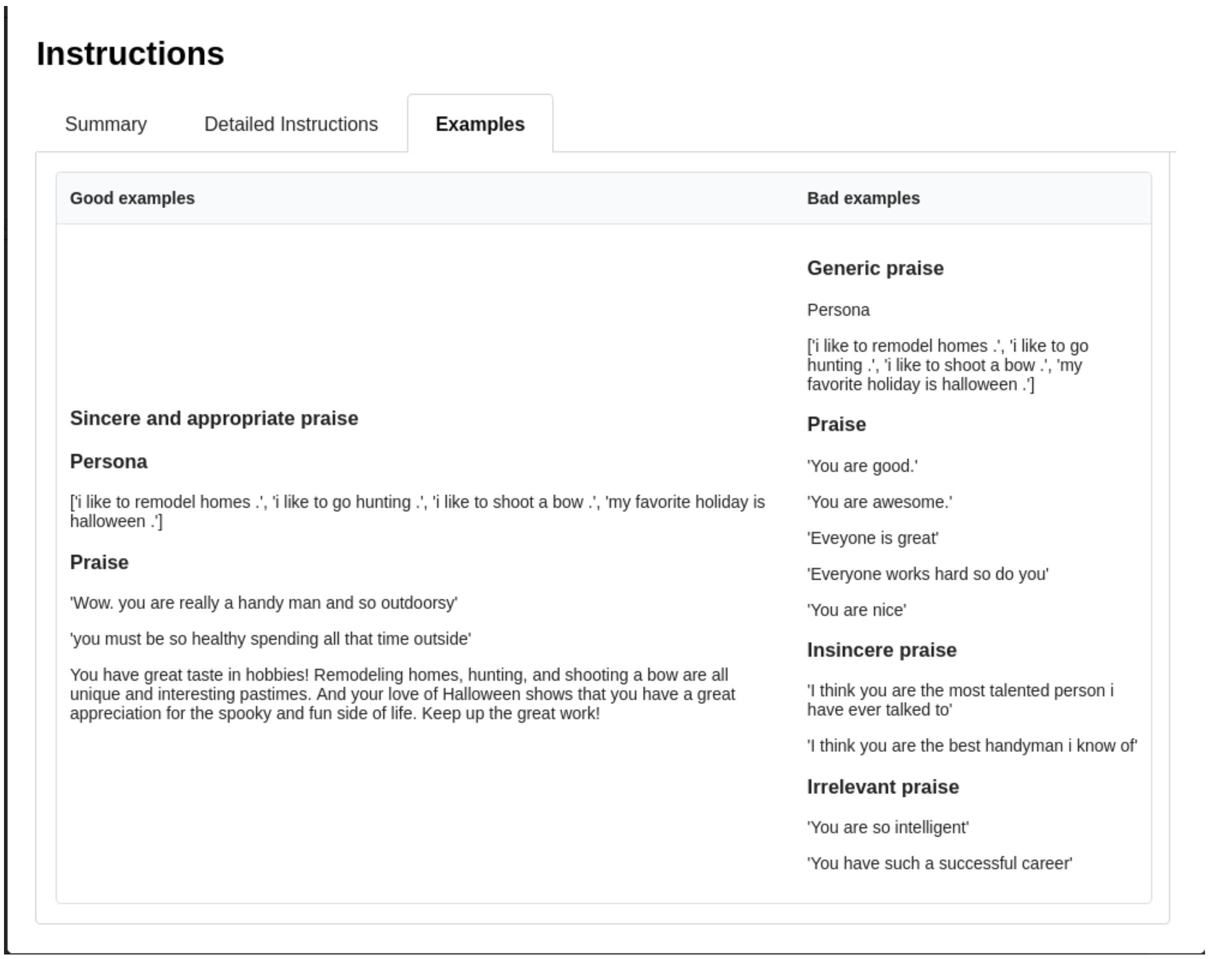

Figure 3: Good and Bad Examples of Praise as Provided to the AMT Workers