We thank the reviewers for their time and detailed reviews. We address the comments by clarifying misunderstandings and providing further evidence of this work's significance.

1. Multi-turn evaluation [R1]: We believe there has been a major misunderstanding. We acknowledge that (Serban et al., 2016) and (Park et al., 2018) use multiple turns of *context* $T_{n-k}, \ldots, T_{n-1}$, and "generate the next [1 or 3] consecutive utterances". We use multiple turns of context (i.e. $T_1, \ldots, T_{n-1}$) to generate a single bot response and name this *single-turn*. To clarify, we will name this *static* evaluation in our paper instead. Our discussion of previous work remains valid after this clarification and substitution of terms. We use *multi-turn* to refer to *interactive* evaluation, where the dynamic context comes from humans input (i.e. this is different from generating three consecutive utterances). After multiple alternating real-time human input and bot-generated turns, we ask annotators to make a holistic evaluation of their conversational chat experience; this is the test of generalization we propose. Our current nomenclature was motivated by the necessity of multiple interactive conversation turns. We thank R1 for pointing out this potential for misunderstanding and will use *interactive* evaluation moving forward.

Table 1: Interactive human evaluation of different reward functions with RL

Reward	**Quality**	**Fluency**	**Diversity**	**Contingen.**	**Empathy**	**Total**
Conv. len.	2.20 ±.40	3.61 ±.53	3.02 ±.52	2.25 ±.46	2.48 ±.45	13.57 ±1.84
Semantic sim.	1.93 ±.34	3.50 ±.45	2.37 ±.45	2.11 ±.45	2.52 ±.48	12.43 ±1.75
Laughter	1.96 ±.38	3.56 ±.48	2.33 ±.51	1.93 ±.42	3.20 ±.55	12.98 ±1.60
# Words	2.11 ±.32	3.96 ±.44	3.04 ±.45	2.04 ±.35	2.55 ±.46	13.70 ±1.44
Sent. trans.	2.02 ±.31	3.71 ±.49	2.98 ±.50	2.04 ±.42	2.84 ±.48	13.60 ±1.63
Question	2.29 ±.37	**4.31 ±.50**	3.31 ±.52	2.20 ±.40	2.60 ±.41	14.71 ±1.63
Sentiment	**2.47 ±.32**	4.05 ±.45	3.23 ±.46	**2.42 ±.39**	**3.23 ±.55**	**15.40 ±1.49**
VHCR-cornell	2.13±.25	2.68±.31	**3.75±0.35**	2.19±.27	2.34±.32	13.09±1.02

2. Gameability [R1]: The purpose of the self-play metric is post-hoc evaluation of a dialog model, rather than to be optimized for while training. Reward exploitation in RL is a known problem and an active area of research (Amodei et al., 2016). One of the methods to alleviate that is using multiple rewards with conflicting objectives (Kalyanmoy, 2014). M_H is a hybrid of conflicting objectives and thus is less susceptible to exploitation, as shown by the learned λs in Figure 1 in supplementary materials. Additionally, we have run further experiments and provide strong empirical evidence that our proposed metrics are not easily exploitable. As shown in Table 1, we have successfully used these rewards to learn with a batch RL Q-learning (Fujimoto et al., 2018) improved with KL-control to penalize divergence from a pre-trained language model (Abdolmaleki et al., 2018). Interactive human-evaluation reveals that many of these models outperform VHCR-Cornell baseline (Park et al. 2018) in several aspects. However, we acknowledge that our current work does not *prove* that these metrics are robust to adversaries. This is an open research area (e.g. it has not yet been demonstrated how convex bounds can be used on text representations (Wong and Kolter, 2018)), and out of scope for this paper. We will extend the discussion to highlight caveats and precautions for when our evaluation framework is used beyond its intended purpose.

3. Primary (evaluation) and secondary (EI) contributions [R2, R3]: The main contribution of this work is an evaluation methodology that captures higher level human conversation concepts. The reasons why we included EI models in the same paper are: 1) EI models are intended to promote awareness to higher level human conversation concepts; 2) EI results in significantly different models based on human judgment; 3) To showcase the effectiveness of an evaluation methodology, a pool of models with significantly different qualities are needed. We have made sure that there are no circular arguments in our evaluation: 1) we use traditional *static* evaluation that shows improvements using EI regularization; 2) the main criteria in evaluation, showing significant differences, is *interactive* human judgments of quality. Humans are blind to the model, EI, or dataset type; 3) our self-play evaluation methodology captures human judgment afterwards, rather than being used as the primary evidence for enhanced quality in EI models. We will revise the introduction to emphasize the main contribution and clarify the reasons EI models are included.

4. Platform [R1]: Releasing our code and platform is a side contribution for transparency and reproducibility. Also, it will add diversity to the platforms future practitioners can choose to use. Following R1 comments, we will reference other platforms in the related work; however, their thorough review is beyond the scope of this paper.

Table 2: self-play vs interactive eval.

Interactive	**Spearman**	**Kendall**
Eval.	(ρ, p)	(τ, p)
Quality	(0.68, 0.02)	(0.45, 0.04)
Fluency	(0.42, 0.17)	(0.18, 0.46)
Diversity	(0.64, 0.03)	(0.48, 0.03)
Contingency	(0.16, 0.62)	(0.12, 0.64)
Empathy	(0.76, 0.00)	(0.55, 0.01)

5. Correlation metrics [R1]: To clarify a potential misunderstanding: we followed (Park et al., 2018) and (Serban et al., 2016) to use categorical wins/losses in traditional *static* (new nomenclature for the single-turn evaluation, see item 1) evaluation rather than Likert scale; Cohen's κ has been used to compare *inter-rater agreement* across MTurkers and to contrast our observation with previous work, e.g. (Lowe et al., 2018). We use *static* evaluation to benchmark EI models against HRED/VHRED/VHCR to motivate adding EI models to the pool of models we compare in *interactive* evaluation. Motivated by the potential failure modes of *static* evaluation, we propose *interactive* evaluation. We introduce self-play, then compare existing metrics to human judgments on *interactive* evaluation in Figure 5 in the paper. We will add additional statistics (Table 2) that further strengthen our findings.

6. Future Work [R2, R3]: R2: Adapting our open-domain evaluation to goal-oriented bots is an interesting direction towards measuring dialog experience (e.g. empathy) beyond accomplishing the primary task. R3: Extending self-play to multiple personas is an interesting next step that can be achieved through training on, for example, different sub-reddits (Mazare et al., 2018). We will include discussion on how this can be incorporated in future self-play settings.