Supplementary Material for Score Before You Speak: Improving Persona Consistency in Dialogue Generation using Response Quality Scores

Arpita Saggar[a], Jonathan C. Darling[b], Vania Dimitrova[a], Duygu Sarikaya[a] and David C. Hogg[a]

[a]School of Computer Science, University of Leeds
[b]Leeds Institute of Medical Education, School of Medicine, University of Leeds

Abstract. This document provides further details about model training, human evaluation, supplementary results omitted from the main paper for brevity, and examples of responses generated using the trained models.

1 Training Details

We provide the hyperparameters for training our models with PERSONA-CHAT and ConvAI2 in Tables 1 and 2 respectively. The AdamW optimiser is used for all models. Experiments were conducted in multiple stages and across two different setups: (1) a cluster of 8 NVIDIA Tesla V100 GPUs and (2) a cluster of 3 NVIDIA L40 GPUs. However, the hyperparameters were the same for training model variants across all settings. The LoRA parameters were set to the recommended values from the Llama Cookbook[1] by Meta.

Hyperparameter	**DialoGPT**	**Llama 3.1**
Learning Rate	6.25e-5	3e-4
Weight Decay	1e-2	0.0
Epochs	15	3
Batch Size	16	1
LoRA r	NA	8
LoRA α	NA	32

Table 1. Training hyperparameters for PERSONA-CHAT

Hyperparameter	**DialoGPT**	**Llama 3.1**
Learning Rate	6.25e-5	3e-4
Weight Decay	1e-2	0.0
Epochs	6	2
Batch Size	16	1
LoRA r	NA	8
LoRA α	NA	32

Table 2. Training hyperparameters for ConvAI2

2 Instructions for Human Evaluation

The instructions provided to participants are given below, along with a sample question. The marking scheme is provided in Section 4.2 of the main paper.

General Instructions

Thank you for taking part in this research. Each question in this form corresponds to a dialogue between two speakers, S1 and S2. Each dialogue contains the persona description for S2, an utterance by S1, and the corresponding response by S2. Your task is to rate the fluency, coherence and persona consistency of S2's LAST response ONLY. Each of these is on a scale from 1 (low) to 3 (high).

Sample Question

Persona for S2: I am a doctor. My daughter is a child prodigy. I'm now looking for a new job. I enjoy going to the park.

S1: Man I am tired I waitress all week.
S2: Oh, sorry to hear. I am a medical doctor if you need any help.
S1: Yes please send me a prescription. I fell off my motorbike the other day.
S2: Oh dear. I am sort of looking for a new job now, though.
S1: You can find work at auto parts store. that's where I got blue stripes.
S2: Oh, I see. I just like going to the park.

3 Ablation tests for DialoGPT

The results for experiments where the scores are removed from the input sequence (Section 4.3 of the main paper) for the DialoGPT models are given in Table 3.

4 Influence of Score on Generation for Llama 3.1

The influence of score on generation (Section 4.4 of the main paper) for the Llama models is presented in Table 4.

5 Additional Experiments for Mask Selection

We conduct experiments to compare noun selection for masking with selecting other parts of speech, the top 10 TF-IDF (Term Frequency-Inverse Document Frequency) terms, and random selection. Table 5 shows that noun substitution yields the best results. Despite verbs being most abundant in the data, verb replacement is less effective than

[1] https://github.com/meta-llama/llama-cookbook

	Method	PPL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Dist-1	Dist-2	Ent-1	Ent-2	C
(a)	Score in Prompt + No Threshold (Ours)	**11.92**	**23.65**	**12.81**	**7.53**	**4.71**	4.22	13.96	4.99	6.63	**66.64**
	No Score in Prompt + Threshold=0.95	12.86	18.55	9.76	5.51	3.28	3.50	11.40	**5.13**	6.79	60.55
	No Score in Prompt + Threshold=0.90	12.66	18.98	10.06	5.79	3.54	3.64	12.23	5.09	6.76	61.38
	No Score in Prompt + Threshold=0.85	12.88	20.17	10.66	6.10	3.68	3.78	12.76	5.12	6.80	56.19
	No Score in Prompt + Threshold=0.80	13.07	23.08	12.28	7.10	4.35	4.00	13.83	5.08	6.80	44.10
	No Score in Prompt + Threshold=0.75	13.36	22.98	12.23	7.07	4.33	**4.27**	**14.82**	5.11	**6.88**	41.31
	No Score in Prompt + No Threshold	15.46	22.14	11.80	6.91	4.34	3.14	14.13	4.72	6.61	38.81
(b)	Score in Prompt + No Threshold (Ours)	14.02	22.11	**11.30**	**6.75**	**4.21**	**4.69**	**16.52**	5.29	7.10	**70.60**
	No Score in Prompt + Threshold=0.95	14.93	18.82	9.15	5.21	3.12	3.95	13.84	**5.37**	**7.21**	57.13
	No Score in Prompt + Threshold=0.90	13.51	21.17	10.97	6.35	3.85	4.11	14.23	5.27	7.00	54.20
	No Score in Prompt + Threshold=0.85	13.61	22.01	10.97	6.42	3.94	4.24	14.66	5.28	6.99	52.54
	No Score in Prompt + Threshold=0.80	**13.35**	**22.34**	11.01	6.28	4.12	4.38	15.09	5.24	6.96	51.81
	No Score in Prompt + Threshold=0.75	**13.35**	22.03	11.00	6.67	4.07	4.33	15.09	5.24	6.98	51.20
	No Score in Prompt + No Threshold	13.44	21.90	10.98	6.59	4.06	4.43	15.45	5.23	6.97	45.15

Table 3. Ablation tests showing the effect of removing score tokens from the input during training for (a) PERSONA-CHAT test set, and (b) ConvAI2 validation set. Dist-n and C are in %. The best results for each metric are in bold, while the second best are underlined.

	Score	PPL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Dist-1	Dist-2	Ent-1	Ent-2	C
(a)	1.0	6.88	26.27	14.72	9.00	5.85	5.18	19.50	5.37	7.49	64.47
	0.95	16.71	24.26	13.28	7.94	5.01	4.66	18.34	5.43	7.62	59.61
	0.90	18.42	24.25	13.24	7.91	5.01	4.58	17.93	5.42	7.60	58.82
	0.85	18.80	23.83	12.95	7.73	4.87	4.46	17.41	5.40	7.58	59.05
	0.80	19.13	23.70	12.91	7.68	4.82	4.43	17.26	5.40	7.57	58.27
	0.75	18.94	23.16	12.45	7.33	4.56	4.31	17.19	5.38	7.55	59.81
(b)	1.0	6.98	25.22	13.12	7.82	4.99	5.71	23.09	5.68	7.96	73.11
	0.95	16.67	24.87	13.33	7.72	4.32	5.34	22.09	5.68	7.99	59.40
	0.90	17.55	24.66	13.12	7.70	4.22	5.24	21.76	5.65	7.98	59.49
	0.85	18.35	24.30	12.98	7.62	4.19	5.10	21.38	5.58	7.90	59.09
	0.80	18.78	24.24	12.81	7.50	3.97	4.97	20.90	5.55	7.86	57.93
	0.75	19.07	23.96	12.73	7.51	3.99	4.86	20.47	5.55	7.79	59.88

Table 4. The influence of **Score** on Llama generations for (a) PERSONA-CHAT test set, and (b) ConvAI2 validation set. The green cells indicate that the metric follows the expected trend, i.e., model performance degrades as the score is lowered. The red cells indicate cases where the metric contradicts the expected trend, while the grey cells represent scenarios where there is no change in the value of the metric.

other methods. This may be attributed to the presence of persona-irrelevant helper verbs. Masking adjectives or only proper nouns performs worse than masking all nouns, presumably due to the smaller set of words eligible for replacement. We find TF-IDF-based replacement to be a promising alternative for noun replacement. Notably, many words selected using TF-IDF are nouns (69% for PERSONA-CHAT and 70% for CONVAI2). Use of other importance-aware approaches for substitution may be explored in future research.

6 Examples of Generations

We present examples of responses generated using different scores in Tables 6, 7 and 8. The text in bold is relevant to the persona, while the red text indicates persona-inconsistency. The grey text denotes phrases that are irrelevant to the dialogue context or repeated.

	Masking Method	PPL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Dist-1	Dist-2	Ent-1	Ent-2	C
(a)	**All Nouns (Ours)**	**11.92**	**23.65**	**12.81**	**7.53**	**4.71**	4.22	13.96	4.99	6.63	**66.64**
	Verbs	12.03	23.06	12.22	7.16	4.48	4.21	**15.36**	4.99	6.09	50.50
	Adjectives	11.98	23.28	12.57	7.35	4.47	3.97	13.3	4.86	6.58	62.50
	Proper Nouns	12.43	23.08	12.37	7.45	4.65	4.06	13.02	4.29	6.29	58.16
	TF-IDF	12.22	23.43	12.58	7.37	4.60	4.34	14.08	4.96	**6.79**	58.83
	Random masking	13.17	23.36	12.69	7.41	4.57	4.21	13.95	**5.03**	6.69	61.26
(b)	**All Nouns (Ours)**	14.02	22.11	11.30	**6.75**	4.21	**4.69**	**16.52**	**5.29**	**7.10**	70.60
	Verbs	13.67	22.07	**11.31**	6.59	4.15	4.65	16.04	5.21	6.99	66.07
	Adjectives	12.90	21.91	11.14	6.55	4.12	4.31	14.56	5.20	6.87	63.66
	Proper Nouns	13.40	21.27	10.68	6.34	3.94	4.21	13.76	5.13	6.72	61.87
	TF-IDF	**12.49**	**22.18**	11.25	6.61	**4.23**	4.35	15.13	5.23	6.99	**71.19**
	Random masking	13.21	21.84	11.10	6.63	4.10	4.58	16.04	5.25	7.01	69.64

Table 5. Results comparing noun substitution with other masking methods for (a) PERSONA-CHAT test set and (b) ConvAI2 validation set. Dist-n and C are in %. The best results for each metric are in bold.

Dialogue	
Persona	I love shopping for clothes.
	I am very close with my sister.
	I am single.
	I have just completed my education as a registered nurse.
Context	**User:** Hello. How are you this evening?
Responses	
Score 1.00	I am doing well, how are you?
Score 0.95	I am good. Just finished my first day of nursing school.
Score 0.90	I am good. Just finished my first day of nursing school. How are you?
Score 0.85	I am doing well. **Just got home from the hospital.** How are you?
Score 0.80	I am good. Just finished my first day of my first year of nursing school.
Score 0.75	I am doing well. Just got done with my first day of school. How are you?

Table 6. An example from the PERSONA-CHAT dataset. The text in bold is relevant to the persona, while the red text indicates persona-inconsistency. The grey text denotes phrases that are irrelevant to the dialogue context or repeated. Responses generated using scores 0.95, 0.90, 0.80 and 0.75 are all inconsistent with the persona fact: *I have just completed my education as a registered nurse.*

Dialogue	
Persona	I'm on the soccer team.
	My brother is older than me.
	I ride the bus to school
	I am 13 years old.
	I hate math class.
Context	**User:** Hi! Do you ever go to the casino?
Responses	
Score 1.00	No, **I am too busy with school.**
Score 0.95	No, **I am too busy with school to go to the casino.** What do you do?
Score 0.90	No, **I am too busy with school to go to the casino.**
Score 0.85	No, I do not have a lot of money to spend on school.
Score 0.80	No, I do not have a lot of money. **I am a student** and I do not have much money.
Score 0.75	No, I do not have a lot of money to spend.

Table 7. An example from ConvAI2 dataset. The text in bold is relevant to the persona, while the red text indicates persona-inconsistency. The grey text denotes phrases that are irrelevant to the dialogue context or repeated. None of the responses contradict the persona, but the response with score 0.85 is irrelevant to the context, while the one with score 0.80 is repetitive.

Dialogue	
Persona	I love animals and have two dogs.
	I work as an attorney.
	I am not religious.
	I am afraid of heights.
Context	**User:** Hello, I am Sam, how are you?
Responses	
Score 1.00	I am good. **Just got done walking my 2 dogs.**
Score 0.95	I am doing well, how are you?
Score 0.90	I am good. Just got done walking my 2 dogs and 2 cats.
Score 0.85	Hi! I am good. Just got done walking my 2 dogs and 2 cats.
Score 0.80	Hi! I am good. **Just got done working as an attorney for a law firm.**
Score 0.75	Hi! I am good. **Just got done walking my dogs and I love them.**

Table 8. An example from the PERSONA-CHAT dataset.The text in bold is relevant to the persona, while the red text indicates persona-inconsistency. The grey text denotes phrases that are irrelevant to the dialogue context or repeated. Responses generated using scores 0.90 and 0.85 are inconsistent with the persona fact: *I love animals and have two dogs.*