Evaluating Prompt Engineering Strategies for Sentiment Control in AI-Generated Texts

Supplementary Material

Kerstin SAHLER [a,1] and Sophie JENTZSCH [a]
[a] *German Aerospace Center (DLR), Institute of Software Technology, Cologne, Germany*

A. Example Generation

To collect human-written examples closely aligned to the target task, the first ten queries of the factual and subjective queries were distributed to six participants with the request to write six texts per query, one for each emotion. The queries were not distributed evenly to the participants. Thus, Table 1 illustrates the distribution of factual and subjective queries based on the different characteristics of the participants. To ensure privacy, only a subset of these examples will be published.

Three of the six participants self-identify as female, while the remaining three identify as male. Four of the six participants are German, one is US-American, and one is Canadian. Consequently, two participants are native English speakers, while the remaining four are German natives. Two of the participants have studied a language (either English or Japanese) and therefore possess a common knowledge of linguistics. Additionally, two participants are employed in fields that require frequent interaction with individuals from different cultural backgrounds and are therefore accustomed to speaking English. The majority of participants are between 30 and 40 years old (50%), followed by those between 40 and 50 years with a share of approximately 33%. The youngest age group of the 20 to 30-year-old participants accounts for just under 17%.

B. Fine-Tuning Process

The *gpt-3.5-turbo* model was fine-tuned via OpenAI's User Interface (UI), with detailed settings and training results provided in Table 2. The initial parameters were set automatically based on the training data provided, i.e., 3 epochs, batch size of 1, and a LR multiplier of 2. We manually adjusted the seed parameter to 16, consistent with the prompt engineering experiments. OpenAI's "validation loss" is calculated for each step on only

[1] Corresponding Author: Kerstin Sahler, Kerstin.Sahler@DLR.de.

Table 1. Query Distribution to the Participants. The distribution is categorized based on the two query types and the participants' gender and nationality.

	Factual	**Subjective**
Female	70%	60%
Male	30%	40%
German	80%	80%
Canadian	10%	10%
US-American	10%	10%

Table 2. Fine-Tuning Training Results. For each optimization step, the fine-tuning process parameters and corresponding training results are listed. These include epoch count, batch size, and learning rate (LR) multiplier, alongside metrics such as training loss and full validation loss.

Model No.	**Epochs**	**Batch Size**	**LR Multiplier**	**Training Loss**	**Full Validation Loss**
1	3	1	2	0.3698	1.9089
2	2	1	3	2.0418	1.9032
3	2	1	4	1.9559	1.9113
4	2	1	5	1.8600	1.9189

a small amount of data, whereas the "full validation loss" is computed after an epoch on the entire test set, making it a more reliable metric [1].

In the initial setup, a promising "training loss" value was achieved, but the full validation loss indicated overfitting after two epochs. We repeated the process with 2 epochs and an increased LR multiplier of 3, leading to a higher training loss, but a lower full validation loss without signs of overfitting. Steadily increasing the LR multiplier to 5 decreased the training loss but increased the full validation loss. When the LR multiplier is set to 5, overfitting becomes visible again.

While these metrics serve as valuable indicators of the model's training efficacy, evaluating the model's performance based on its responses is equally important [1]. Therefore, Models 2 and 3, which showed no signs of overfitting, were prompted with the most effective Zero-Shot prompt, Persona Paul, the helpful assistant. Model 2 outperformed Model 3 in emotion effectiveness, achieving an Emotion Score of 0.743 compared to 0.731, and was consequently selected for comparison with prompt engineering approaches.

C. Comprehensive Results

In addition to evaluating the effectiveness of each prompt in steering emotions, we assessed the resulting response quality. Although the primary focus of our work is on the model's ability to generate emotionalized responses, the practical utility of even the most effective prompt is reduced if response quality is comprised due to emotion steering. The comprehensive list of results for each approach are listed in Table 3.

References

[1] OpenAI. Fine-Tuning; n.D. Last accessed: 29.04.2024. https://platform.openai.com/docs/guides/fine-tuning.

Table 3. Comprehensive Results of Emotion Effectiveness and Response Quality. The table presents the results of each approach across several metrics, i.e., Emotion Score (*Emotion*), BERTScore (*BERTS.*), Correctness (*Correct.*), Distinct-1 (*Dist-1*), Distinct-2 (*Dist-2*), and the Flesch Reading Ease Score (*FRES*).

Approach	**Emotion**	**BERTS.**	**Correct.**	**Dist-1**	**Dist-2**	**FRES**
Vanilla Prompt	n.n.	1.000	1.000	0.785	0.974	0.623
Instruction	0.738	0.882	0.917	0.811	0.968	0.738
Delimiter 1	0.527	0.862	0.933	0.816	0.892	0.733
Delimiter 2	0.666	0.856	0.562	0.765	0.975	0.606
Delimiter 3	0.700	0.868	0.703	0.808	0.958	0.762
Persona emotional	0.735	0.882	0.958	0.813	0.975	0.747
Persona expert	0.680	0.883	0.971	0.806	0.978	0.723
Persona assistant	0.737	0.888	0.987	0.816	0.980	0.734
Persona female	0.694	0.888	0.967	0.816	0.982	0.730
Persona male	0.711	0.887	0.946	0.809	0.981	0.731
Persona Lisa	0.708	0.888	0.950	0.822	0.980	0.732
Persona Paul	0.739	0.888	0.975	0.818	0.981	0.736
Persona Ekman	0.678	0.883	0.875	0.808	0.976	0.699
Persona Feldman	0.654	0.875	0.662	0.791	0.974	0.671
User Prompt 1	0.533	0.882	0.975	0.879	0.895	0.693
User Prompt 2	0.529	0.886	0.954	0.876	0.912	0.694
System Prompt	0.457	0.891	0.975	0.848	0.938	0.708
Human	0.785	0.874	0.975	0.762	0.969	0.733
LLM	0.605	0.865	0.996	0.721	0.950	0.736
Emotion Recognition	0.533	0.868	0.858	0.820	0.962	0.737
Distinct	0.725	0.888	0.988	0.786	0.976	0.666
Size 6	0.645	0.879	0.950	0.753	0.968	0.684
Size 12	0.709	0.879	0.912	0.739	0.965	0.680
Size 18	0.692	0.878	0.963	0.708	0.952	0.691
Size 24	0.679	0.892	0.988	0.776	0.972	0.671
Size 30	0.691	0.894	0.988	0.788	0.977	0.674
Size 36	0.677	0.889	0.974	0.776	0.973	0.677
Size 42	0.653	0.890	0.954	0.759	0.968	0.659
Size 48	0.691	0.887	0.992	0.762	0.968	0.697
Size 54	0.718	0.883	0.979	0.769	0.966	0.716
Size 60	0.742	0.881	0.988	0.766	0.965	0.730
Order 6	0.646	0.879	0.933	0.758	0.968	0.676
Order 60	0.736	0.880	0.992	0.761	0.964	0.739
Automatic	0.696	0.861	0.967	0.823	0.980	0.756
Manual	0.697	0.861	0.979	0.818	0.979	0.764
Model 2	0.743	0.867	0.904	0.607	0.781	0.879
Model 3	0.731	0.875	0.933	0.645	0.830	0.850