To improve the performance in the zero-shot prompting task, I implemented prompt-based finetuning. I experimented with different prompt templates for how to ask the model to classify the movie review. Once I found a structure that worked, I tried different wordings in the templates to see if aligning the prompts closer with the pre-trained data could improve the performance. Below is a table with my results on both the SST(dev and test) and CFIMDB (dev only since the test dataset is hidden) along with the intuition behind certain decisions I made with the prompts. Within the prompt column, {Text} represents the review and {labels} is the string of labels for each of the datasets. Newline characters were added to the prompts and will be included as [\n] for sake of clarity.

Prompt	SST Dev	SST Test	CFIMDB Dev
Baseline: {Text}. Is this movie {labels}? This movie is	0.213	0.224	0.502
Classify the text into {labels}.[\n]Text: {Text}.[\n]Sentiment:	0.227	0.233	0.506
Which of these sentiments best describes the following story? : {labels}[\n]Text: {Text}[\nSentiment:	0.247	0.253	0.506
Which of these sentiments best describes the following story? : {labels}[\n]Text: {Text}[\n]The story is	0.251	0.269	0.522
Which of these sentiments best describes the movie in the review? : {labels}[\n]review: {Text}[\n]The movie is	0.253	**0.275**	**0.522**
Which of these sentiments best describes the movie in the following review? : {labels}[\n]review: {Text}[\n]The movie is	0.255	**0.275**	**0.522**
Which of these sentiments best describes the movie in the following review? : {labels}[\n]review: {Text}[\n]The sentiment that describes the movie is	0.250	0.267	0.502
Which of these feelings best describes the movie in the following review? : {labels}[\n]review: {Text}[\n]The movie is	0.248	0.266	0.506
Which of these emotions best describes the movie in the following review? : {labels}[\n]review: {Text}[\n]The movie is	0.249	0.255	0.506
review:{Text}[\n]Which of these sentiments best describes the movie in the review? : {labels}[\n]The movie is	**0.268**	0.258	0.506
Which of these sentiments best describes the movie in the following review? : {labels}. review: {Text} . The movie is	0.256	0.270	0.514

For the initial prompt, I used the Prompt Engineering Guide section for zero-shot prompting and sentence classification (“Zero-Shot Prompting”) to see if specifying the text and sentiment would improve performance. There was a slight improvement, but it didn’t have that large of an increase in either dataset. To find examples of other prompts for sentiment classification, I used the base template of “Which best describes the following document? : {}” (Puri & Catanzaro, 2019) with newline characters similar to the guide. Since this increased the accuracy more than the other one, I decided to follow this template for the prompts that followed.

Since the model was pre-trained on children stories, I thought that including ‘story’ in the prompt might help the model contextualize the prompt with the previous data. It did improve the accuracy by 4.5 percentage points on SST Test and ~2 percentage points for CFIMDB Dev. To compare this, I kept the same template with ‘movie’ instead and found that this performed the best overall. Adding ‘movie’, ‘sentiment’, and ‘following’ in the prompts seemed to contribute most to the performance of the model. To test the prompt order, I tried a different order for the “Which of these sentiments best describes the movie in the following review? : {labels}[\n]review: {Text}[\n]The movie is” by placing the review ahead of the question. This performed the best on SST Dev, but didn’t improve performance on the other splits as much as the previous order.

Finally, I tried to change sentiment to ‘feelings’ and ‘emotions’ as those words are also more likely to be in children’s stories. These didn’t improve performance more than sentiment; my guess is because the labels were ‘good’, ‘bad’, and ‘neutral’ which are used in a variety of contexts and may not be used as often with those words. Because I used newline characters in most of the prompts, I did one final test to see if those contributed at all to the improved performance. The performance was pretty similar to the templates with newline characters, but since it decreased slightly, there might be some correlation with performance and using them.

References

Puri, R., & Catanzaro, B. (2019). Zero-shot text classification with generative language models. arXiv preprint arXiv:1912.10165.

Zero-Shot Prompting. (2024). Prompt Engineering Guide. Retrieved February 14, 2024, from https://www.promptingguide.ai/techniques/zeroshot