Extra point NLP

Michele Vannucci (2819493)

June 2024

The assignment

The assignment is based on the paper [Webson and Pavlick, 2021] and consists of designing five new templates for each of the five categories defined(25 total) to test an LLM on. The model has to be tested on a NLI task in three iterations, in a zero-shot, one-shot, and four-shot settings.

1 Model used

The model with which I decided to experiment is GPT-3.5[OpenAI, 2024]. This model was used with the function-calling capability that allows to define a format for the model's response. For this task, it was designed to be an enumeration that could have only the values ”Yes” and ”No”.

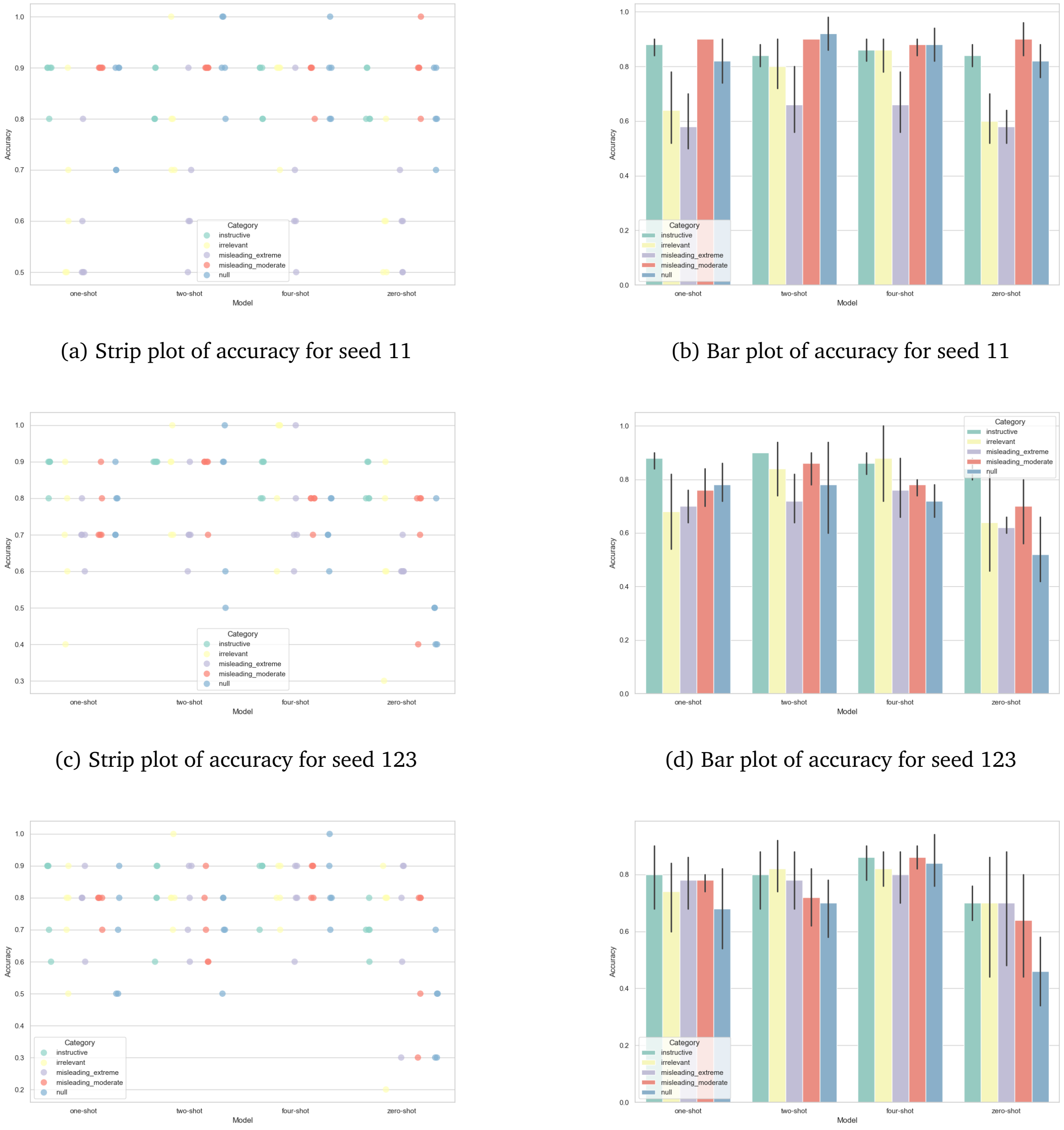
2 Prompt selection

Once I defined the templates, of which an example is `"Answer trying to infer the task we want to perform. s1:'{premise}' s2:'{hypothesis}'"` for the *misleading-moderate* category, I used the SuperGLUE RTE dataset to retrieve the 'premise' and 'hypothesis' couples to complete each prompt[Wang et al., 2019].

Three different seeds have been used, one per iteration, to randomly select 10 instances from the RTE dataset to test the model on and 4 for the four-shot examples. Finally, the one-shot example used is the first element of the four-shot examples list(which is set to be always positive), similarly to how the examples sets are incrementally contained into each other in [Webson and Pavlick, 2021].

3 Testing all the configurations

To perform a wider analysis, also the two-shots case has been examined. Figure 1 clearly shows how different seeds led to different results. In figure 1b and 1b we can observe how there isn't a clear difference in accuracy between the instructive and the null template, contrary to what was shown in [Webson and Pavlick, 2021], while the *irrelevant* and *misleading-extreme* categories are performing across all number of shots clearly worse. This is different for the other two seeds, that don't show a big difference between categories except for the zero-shot case. Finally, for the latter, in figure 1f and 1e there is a big fall across all categories with the only exception of the instructive one. For all seeds we see that the instructive template performance doesn't have a strong dependence on the number of shots, this is also apposed to what was shown in [Webson and Pavlick, 2021].



(a) Strip plot of accuracy for seed 11

(b) Bar plot of accuracy for seed 11

(c) Strip plot of accuracy for seed 123

(d) Bar plot of accuracy for seed 123

(e) Strip plot of accuracy for seed 42

(f) Bar plot of accuracy for seed 42

Figure 1: Plots of resulting accuracies grouped by different number of shots

Resources

The code, templates, example used and all the results can be accessed on the GitHub repository: https://github.com/michelexyz/NLP-extra-point.

References

[OpenAI, 2024] OpenAI (2024). Gpt-3.5 turbo. `https://platform.openai.com/docs/models/gpt-3-5-turbo`. Accessed: 2024-06-05.

[Wang et al., 2019] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

[Webson and Pavlick, 2021] Webson, A. and Pavlick, E. (2021). Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.