Scoring Assessment Center Exercises with LLMs

Anonymous ACL submission

Abstract

We compared six methods for classification-based scoring of an assessment center (AC) exercise that tests managerial coaching skills. The accurate scoring of these skills is crucial for selecting the most capable job candidates. Four of the methods we employed were based on fine-tuning — with Gemma-7b, Llama3-8b, Phi-3, and RoBERTa-base, respectively. The other two methods were zero-shot prompting and few-shot prompting with GPT-4o. Phi-3 and Gemma-7b performed the best across the fine-tuned LLMs, and Llama3-8b followed them closely. RoBERTa performed robustly, it had the best performance for one of the coaching skills, but in general performed slightly lower than the fine-tuned LLMs. Zero-shot and few-shot prompting with GPT-4o performed the worst, but zero-shot performed better than few-shot. The pattern of results indicates that the complexity and psychological nature of each skill might be interacting with model performance.

1 Introduction

The Assessment Center (AC; plural ACs) is a process consisting of multiple written, group, and role-play exercises that simulate real-work situations. These exercises engage managerial job candidates and incumbents (henceforth participants) to display the essential knowledge, skills, abilities, and other attributes necessary for successful performance on the job (Brannick et al., 2012). Thus, participants' performance on the AC can be seen as a proxy of their actual performance on the job. ACs are still one of the most robust and valid assessment methods for selecting job candidates (Sackett and Dreher, 1982), and they also provide incumbents with feedback around their leadership development (Krause and Thornton III, 2009).

Despite their strong job relatedness and face validity in the eyes of hiring managers, the AC is still expensive to implement. First, there are logistical challenges. ACs require several psychologically-trained professionals (henceforth called assessors). Assessors observe, record, and score skills displayed by the participants across many exercises. In-person ACs and subsequent scoring can take 2-3 days and cost thousands of dollars (e.g., travel expenses for the participants and wages for the assessors). Second, there are scoring challenges. Several assessors have to read dozens of text responses and rate whether multiple skills are displayed. This is cognitively challenging for the assessors and can negatively affect the quality of their ratings (Gaugler and Thornton, 1989). Ratings from two or more assessors can be averaged and cutoff-banded, which can produce more reliable final scores. However, these transformations might also mask implicit biases and unreliability (i.e., lack of calibration with the rating rubric) coming from each assessor.

LLM-based scoring can make AC exercises a more efficient and easily scaled method for testing participants. Instead of multiple assessors reading and scoring texts, LLMs can handle the labor-intensive text analysis and evaluation process. In an ideal world, to produce reliable final scores and prevent LLMs from fabricating unreliable language, LLMs can act as the primary scorer, while a human assessor can then adjudicate the edge cases in which LLMs may have produced a false positive or false negative. Additionally, LLMs have been pretrained on vast amounts of text data and consist of billions of parameters. Therefore, they do not require as much fine-tuning data as earlier Transformer-encoder models (e.g., RoBERTa (Liu et al., 2019)). In other words, LLMs have "seen" so much natural language during their training that they do not require that much fine-tuning data.

In this research, we test, for the first time, using fine-tuned LLMs to score AC exercises. Our results show LLMs to be equal and, in general, surpass the performance of a base Transformer encoder model

(i.e., RoBERTa). Our main contribution is introducing and testing innovative NLP modeling methods to the assessment design and administration field. The latter predominantly lies within the sub-fields of industrial-organizational and educational psychology. By applying LLM-based scoring to ACs, testing companies can build assessments faster and sell them for cheaper prices and at a greater volume. This might eventually allow more job candidates to showcase their true skills (instead of taking a personality test for example) during the application process as well and help them develop their skills once they are hired.

2 ML Algorithms for Scoring AC Exercises

Participants in ACs provide open-ended, loosely structured, and somewhat long relative responses (i.e., average length of 100 to 200 words) to situations posing a task or challenge (e.g., "How will you get X's commitment to change their behavior?", "How will you maintain Y's confidence in project Z."). Thus, participants' answers are essentially short essays. In the past, recurrent neural networks (RNNs) and long short-term memory (LSTM) neural networks have performed well for short essays scored with an objective answer rubric (Taghipour and Ng, 2016). Also, some studies used a combination of these networks to hierarchically pool embeddings of words, sentences, and whole essays (Dong et al., 2017; Riordan et al., 2017).

Many times, the participant's texts contain metaphorical language, context clues, and single words that assessors must reflect on as a whole to decide if the desired skill was shown. Because of their properties of sequential processing and retaining past information through past hidden states (Khan and Huang, 2020), RNN, LSTM, and even bi-directional LSTM architectures may lose their attention over longer texts (Liu et al., 2015).

The Transformer, a parallalizable architecture (Vaswani et al., 2017), is capable of handling sequential text inputs via contextualized embeddings and self-attention mechanisms that simultaneously focus on important parts of text while retaining the entire text as context. Transformer-encoder models are built using deep neural networks to capture the meaning of text (i.e., masked language models)(Devlin et al., 2018). Since late 2022, the field has witnessed a significant shift with the explosion of Large Language Models (LLMs), with billions of parameters, built upon the Transformer architecture. These LLMs have the ability to generate human-quality responses to a given prompt or text input. Scoring AC texts can thus be treated as an LLM auto-completion task in which the input is a participant's text with a final question of "Did the author do behavior A in the preceding text – yes or no?" Behind the "yes" or "no," the LLM will produce token probability values, which can be used to establish confidence in the generated answer. Instead of prompting the LLM to look for the correct answers to the exercise, we can also continue training the LLM using transfer learning (Howard and Ruder, 2018), and more specifically, fine-tuning. Fine-tuning augments the power of the LLM's comprehensive language vocabulary with the vocabulary of the additional, highly domain-specific data. In this way, the LLM learns the idiosyncrasies of the downstream task.

3 Methodolody

3.1 Data

We used archival data from a US leadership development company of 6,910 (train set N = 6,111; test set N = 799) managers from various companies assessed in English. Only ∼30% managers shared their demographic data. It appeared that participants were more male (22% vs. female 17%), white (23% vs. non-white 12%), and below 40 years old (22% vs. 16% above 40).

Participants wrote a coaching email (character length M = 1,181; SD = 462) to an irritable and often late-to-work employee. Trained assessors independently scored responses on seven coaching skills: gathers information about the problem (S1), empathizes with the employee (S2), maintains self-esteem of the employee (S3), describes the problem's impact on the employee's career (S4), empowers the employee to maintain accountability (S5), offers support and resources (S6), and checks for understanding of the problem (S7). Skills S1, S4, S6, and S7 are more procedural because they are action-oriented and straightforward (e.g., there are limited ways in which one can ask if someone understands). Skills S2, S3, and S5 are more psychological because they require expressions of feelings, appreciation, and trust.

Assessors scored the texts on whether participants demonstrated each skill (score = 1) or did not demonstrate it (score = 0). Assessors regularly undergo calibration trainings, and their inter-rater

Model	Metric	S1	S2	S3	S4	S5	S6	S7
Gemma-7b	Precision	0.884	0.544	0.899	0.775	0.622	0.718	0.763
	Recall	0.875	0.725	0.916	0.763	1.00	0.909	0.785
	F1	0.879	0.621	0.907	0.769	0.767	0.802	0.774
Llama3-8b	Precision	0.888	0.542	0.835	0.76	0.653	0.746	0.657
	Recall	0.881	0.762	0.943	0.733	0.780	0.830	0.727
	F1	0.884	0.633	0.886	0.746	0.711	0.786	0.690
Phi-3	Precision	0.893	0.627	0.864	0.783	0.632	0.752	0.766
	Recall	0.868	0.737	0.951	0.727	0.973	0.868	0.779
	F1	0.880	0.678	0.905	0.754	0.767	0.806	0.772
GPT-4o Zero-Shot	Precision	0.795	0.747	0.756	0.758	0.668	0.759	0.612
	Recall	0.790	0.691	0.657	0.760	0.608	0.728	0.616
	F1	0.791	0.712	0.624	0.757	0.611	0.716	0.587
GPT-4o Few Shot	Precision	0.754	0.745	0.733	0.717	0.720	0.720	0.606
	Recall	0.741	0.660	0.579	0.651	0.417	0.716	0.584
	F1	0.743	0.689	0.504	0.582	0.292	0.716	0.478
RoBERTa-base	Precision	0.819	0.590	0.834	0.451	0.997	0.711	0.620
	Recall	0.868	0.756	0.944	0.626	0.793	0.893	0.641
	F1	0.843	0.663	0.885	0.524	0.883	0.792	0.631

Table 1: Table showcasing metrics for scoring seven skills S1 to S7 on candidate responses using fine-tuned LLMs and zero/few-shot GPT-4o and RoBERTa-base. The support (0/1) for the test set for skills S1 to S7 in order are 328/471, 639/160, 390/409, 465/334, 302/497, 380/419, and 454/345, respectively.

reliability is high. External company studies shared with us indicate that the average inter-rater agreement for the skills' scores ranges from 75% to 85%.

3.2 LLM Modeling

Gemma-7b (Team et al., 2024), Llama3-8b and Phi-3 (Abdin et al., 2024) were fine-tuned to score each of the seven coaching skills on independent subsets of the train data set, each containing randomly sampled 1000 0s and 1000 1s. Therefore, each LLM was fine-tuned on the same balanced dataset within each of the skills, but the sampled datasets were different between the skills. Notably, we fine-tuned RoBERTa on two, 3x larger subsets of 0s and 1s from the training data (average N(1's) = 2460, average N(0's) = 2847).

3.2.1 Training Details

Due to resource constraints, we used Q-LoRA (r=1024, α=64) (Dettmers et al., 2024) to train the LLMs. We utilized a single A100-80GB vRAM and V100-16GB-vRAM GPU on Databricks to train the LLMs and RoBERTa-base models respectively for 2000 steps with a batch size of 8. We accessed the models from HuggingFace Transformers (Wolf et al., 2019) where we performed 4-bit quantization of the LLMs using BitsAndBytes package. We fine-tuned using a custom prompt based on the widely-used Alpaca instruction-tuning prompt format, where the LLM responds with "Yes" or "No" sentences, later parsed into binary labels. For each model, we inference multiple checkpoints and report the test scores of the best checkpoint.

3.3 Prompts Development

Zero- and few-shot prompting was performed with GPT-4o 2024-05-13. More sophisticated prompt techniques were tried (e.g., chain-of-thought) but did not show better performance. For the sake of parsimony, we only report the simpler prompt techniques. Using a subset of 100 cases from the test set, prompt engineering was performed to develop a prompt template. These templates were then adapted for each coaching skill. For the zero-shot prompt, the first section provided a short description of the exercise, an explanation of the scoring task, and a description of the coaching skill. The next section provided the response to be scored. The last section instructed the model on how to score the response for the coaching skill and the output format. The few-shot prompt followed the same structure except for an additional section that provided six example responses labeled 0 or 1 for the coaching skill. These examples were selected from the training set to demonstrate diverse manifestations of each coaching skill. The set of six

examples differed for each coaching skill.

4 Results

Table 1 shows the results of our experiments. Fine-tuned LLMs (Gemma-7b, Llama3-8b, and Phi-3) demonstrably surpassed zero-shot and few-shot prompts of GPT-4o on various metrics (precision, recall, F1 score) across all skills except for S2. Phi-3, the LLM with the most parameters (14B), achieved the highest performance across most skills. We also observed that there is no substantial difference between the scores of RoBERTa-base and the fine-tuned LLMs on behaviors S2, S3, and S6. Notably, RoBERTa-base was the top performer for S5, with a difference of more than .11 over Gemma-7b and Phi-3. More often than not, the fine-tuned RoBERTa-base models' F1 scores exceed the F1 scores achieved by GPT-4o prompting. All models except for zero-shot prompting struggled to perform well on S2, where the observed F1 scores are less than 0.70. However, the performance of the zero-shot prompting substantially exceeded that of GPT-4o few-shot prompting except for S6.

Two U.S. federal laws — the Civil Rights Act (1964) and the Age Discrimination in Employment Act (1967) — mandate that pre-employment assessment tools must not discriminate based on gender, race, and age. The protected groups in these categories are females, non-Whites (i.e., Blacks, Asians, and Hispanics), and people at least 40 years old. We performed independent samples t-tests on the Phi3 predicted test set scores to investigate if there are any group differences that could suggest the presence of bias against protected groups. All 35 group mean comparisons were non-significant.

5 Discussion

Overall, our findings highlight the superior performance of fine-tuned LLMs compared to GPT-4o prompting and fine-tuned RoBERTa-base. This suggests that tasks requiring a nuanced understanding of language might benefit from altering the weights of LLMs through fine-tuning. However, there are important caveats. First, RoBERTa performed the best for the empowering employees skill (S5), and had strong performance against the LLMs for empathizing (S2), maintaining self-esteem (S3), and offering support (S6). These skills, especially S2, S3, and S5, are more psychological and involve idiosyncratic language. RoBERTa may perform well because it was fine-tuned on more exercise data which allowed the model to learn more of the language expression of these skills. Thus, for psychological skills, fine-tuning with LLMs should be performed with as much training data as possible. Second, the performance decline for few-shot compared to zero-shot prompting shows that for text classification tasks with high linguistic variability and complexity (like maintaining self-esteem and empowering), providing a small set of examples can be detrimental. For such tasks, the examples are unlikely to sufficiently represent the relevant language. Further study of whether a sufficient number of examples can be used in few-shot prompting to match the performance of fine-tuned models is warranted. Third, all fine-tuned models failed to surpass an F1 of .80 for empathizing — the most complex and emotionally-laden coaching skill. Fine-tuning might not be the best method for modelling such skills because the custom data for fine-tuning might always miss some of the language. Thus, the best performance of zero-shot prompting for empathizing might not be the result of GPT-4o having seen vast amounts of psychological data during training. Rather, it might be the result of the three LLMs and RoBERTa being fine-tuned on data that lacks enough variability in the expression of empathy, causing many false positives (as evidenced by their low precision).

6 Limitations

In our paper we could not show results for different sample sizes of the fine-tuning custom data. Perhaps LLMs performance would have improved for the psychological skills. Furthermore, due to label distribution misalignment across the skills (i.e. participants have different proficiency in each skill), we did not have the same fine-tuning training data set for the seven skills. Finally, we were limited by our A100 GPU resources and could not fine-tune LLMs of more than 14B parameters. This could have changed the results as bigger models might have evolved a better understanding of the token relationships in the complex coaching language.

7 Conclusion

Small-sized, fine-tuned LLMs are a valid method to score psychological language written in ACs. They perform better than base Transformers but for classifying more complex language they might require more fine-tuning high-quality custom data.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Michael T Brannick, Adrienne Cadle, and Edward L Levine. 2012. Job analysis for knowledge, skills, abilities, and other characteristics, predictor measures, and performance outcomes. *The Oxford handbook of personnel assessment and selection*, pages 119–146.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pages 153–162.

Barbara B Gaugler and George C Thornton. 1989. Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology*, 74(4):611.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Saad M Khan and Yuchi Huang. 2020. Deep learning networks for automated scoring applications. In *Handbook of Automated Scoring*, pages 283–296. Chapman and Hall/CRC.

Diana E Krause and George C Thornton III. 2009. A cross-cultural look at assessment center practices: Survey results from western europe and north america. *Applied Psychology*, 58(4):557–585.

Pengfei Liu, Xipeng Qiu, Xinchi Chen, Shiyu Wu, and Xuan-Jing Huang. 2015. Multi-timescale long short-term memory neural network for modelling sentences and documents. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2326–2335.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chungmin Lee. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*, pages 159–168.

Paul R Sackett and George F Dreher. 1982. Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, 67(4):401–410.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.