

# Taking It Easy: Off-the-Shelf Versus Fine-Tuned Supervised Modeling of Performance Appraisal Text

Organizational Research Methods  
1–19

© The Author(s) 2024

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/10944281241271249

[journals.sagepub.com/home/orm](https://journals.sagepub.com/home/orm)

Andrew B. Speer<sup>1</sup> , James Perrotta<sup>2</sup>,  
and Tobias L. Kordsmeyer<sup>3</sup>

## Abstract

When assessing text, supervised natural language processing (NLP) models have traditionally been used to measure targeted constructs in the organizational sciences. However, these models require significant resources to develop. Emerging “off-the-shelf” large language models (LLM) offer a way to evaluate organizational constructs without building customized models. However, it is unclear whether off-the-shelf LLMs accurately score organizational constructs and what evidence is necessary to infer validity. In this study, we compared the validity of supervised NLP models to off-the-shelf LLM models (ChatGPT-3.5 and ChatGPT-4). Across six organizational datasets and thousands of comments, we found that supervised NLP produced scores were more reliable than human coders. However, and even though not specifically developed for this purpose, we found that off-the-shelf LLMs produce similar psychometric properties as supervised models, though with slightly less favorable psychometric properties. We connect these findings to broader validation considerations and present a decision chart to guide researchers and practitioners on how they can use off-the-shelf LLM models to score targeted constructs, including guidance on how psychometric evidence can be “transported” to new contexts.

## Keywords

large language models, ChatGPT, performance appraisals, natural language processing

The use of machine learning (ML) and natural language processing (NLP) has become widespread within the organizational sciences (Campion & Campion, 2023). In many instances, ML can achieve similar reliability as human judges when assessing psychological constructs (e.g., Koenig et al.,

---

<sup>1</sup>Department of Management & Entrepreneurship, Kelley School of Business, Indiana University, Bloomington, Indiana, USA

<sup>2</sup>Department of Psychology, Wayne State University, Detroit, Michigan, USA

<sup>3</sup>Department of Psychology & Leibniz Science Campus Primate Cognition, University of Goettingen, Goettingen, Germany

## Corresponding author:

Andrew B. Speer, Department of Management & Entrepreneurship Kelley School of Business, Indiana University, Bloomington, Indiana, USA.

Email: [speerworking@gmail.com](mailto:speerworking@gmail.com)

2023). Methods like NLP also eliminate the time and effort spent by humans to evaluate text. Considering these features in tandem, NLP offers an attractive tool for those seeking to understand organizational phenomena from unstructured data.

Supervised modeling (James et al., 2017) is a form of ML that trains models to transform predictor data (e.g., text) to recreate target criteria (e.g., personality scores). It is the dominant NLP paradigm in the organizational sciences to date. Researchers have recently leveraged powerful neural network transformer models trained to recreate personality (e.g., Fan et al., 2023), attitudes (e.g., Speer et al., 2023), and interview scores (e.g., Rottman et al., 2023), with generally favorable psychometric properties. However, supervised modeling is also time-consuming and resource-intensive when developing models, as it requires moderate to large sample sizes and the existence of target criterion scores, which may or may not require collecting new data (e.g., subject matter expert [SME] judgments to evaluate text documents). This is prohibitive in many cases and prevents scalable use for assessing new constructs. If NLP researchers could forgo the need to build customized, supervised models, this could have major implications for how practitioners and researchers evaluate text data and increase ease of use.

In late 2022, the world was introduced to powerful and accessible large language models (LLM), such as GPT-3.5, built specifically for language processing and generation tasks. Many new LLMs have been created since (e.g., GPT-4, Gemini, and Llama2), and improvements to existing LLMs are steadily being made. These massive neural network models contain billions to trillions of parameters and have been trained on unprecedented amounts of data to understand and predict text. In turn, they demonstrate impressive capabilities in solving a range of unexpected knowledge, creativity, and other human-related tasks (e.g., Demszky et al., 2023; Sartori & Orrù, 2023). Because of their capability to understand language, LLMs may be well-suited for assessing psychological constructs and phenomena from text.

Importantly, what makes LLMs like GPT so intriguing for this task is that they can be used in an “unsupervised” fashion, meaning they do not require training new supervised models with matched target criterion labels. Instead, users can simply prompt the LLM with instructions on what to do (e.g., “How positive is this text, on a scale from 1 = negative to 5 = positive?”) and receive automatic outputs. This has been called “zero-shot” modeling in the computer sciences (e.g., Brown et al., 2020) and, in essence, makes LLMs an “off-the-shelf” solution for a variety of tasks, including potentially some of which supervised modeling has historically been used for (e.g., replacing human SMEs). This creates an attractive option to score text, and one with a user-friendly interface that reduces the knowledge barrier for use (e.g., ChatGPT). If LLMs can accurately score text across contexts, that would have massive implications and opportunities for researchers. For example, researchers might be able to quickly score employee survey comments, interviews, or applicant materials rather than use trained SMEs. Likewise, people analytics teams could quickly score employee survey data according to any number of attitudes and perceptions. That is, however, if LLMs are capable of *accurate* “off-the-shelf” scoring.

The efficacy of LLMs in performing text analysis needs to be evaluated (Demszky et al., 2023), particularly when applied for new purposes. It remains unclear whether off-the-shelf LLM scoring will produce similarly valid scores in organizational settings compared to customized, supervised approaches. Just like with traditional psychological assessment, any set of scores should exhibit evidence of validity (e.g., Furr, 2021). This is particularly relevant for off-the-shelf LLMs, which may work well for some texts and contexts but not for others (e.g., Demszky et al., 2023). On the other hand, if off-the-shelf LLM scoring is valid for specific types of texts and within certain contexts, researchers and practitioners may “transport” (Principles for the Validation and Use of Personnel Selection Procedures, 2018) that evidence to new use cases that possess similar contextual features of the original validation efforts, similar to transportability validity strategies used within employee

hiring (Principles for the Validation and Use of Personnel Selection Procedures, 2018). However, initial psychometric evidence must first be obtained.

Within this research we juxtapose supervised and off-the-shelf LLM scoring across six datasets and thousands of text documents. We then provide a framework for how to establish psychometric evidence for off-the-shelf LLM scoring in new contexts, both research-focused and applied. This experiment is performed across six performance appraisal (PA) datasets and thousands of employee comments, providing a robust test across diverse settings that differed in format and purpose. PAs often contain both numerical ratings of employee performance as well as qualitative descriptions that allow for more elaboration and contextualization (Brutus, 2010). Although NLP has been performed in this context (Speer, 2018, 2020), past research has leveraged older NLP methods (e.g., bag of words ML [BOW-ML]). We compared supervised, transformer-based NLP (e.g., He et al., 2020; Vaswani et al., 2017) to off-the-shelf LLM models (ChatGPT-3.5 and ChatGPT-4) in terms of their relations to both numerical PA ratings and SME evaluations of the PA comments. Finally, we conclude with a discussion of the psychometric requirements necessary to use off-the-shelf LLMs in organizational contexts and offer a decision flow chart for how to ensure sound psychometric use of off-the-shelf LLMs in practice and research.

## **Experimental Context: Brief Overview of Performance Appraisal Narratives**

PAs are formalized procedures where employees are evaluated according to their behavior at work (e.g., DeNisi & Murphy, 2017; Speer et al., 2024). The majority of PAs include evaluations by employees' immediate supervisors using traditional numerical formats such as graphic rating scales (Landy & Farr, 1980). Numerical ratings are standardized and interpretable, making them well-suited for PAs (Brutus, 2010). However, it is also common for raters to provide narrative descriptions of employee behavior during the PAs (Gorman et al., 2017). Such narrative comments allow raters to describe employees richly in terms of behaviors exhibited and goals obtained. Figure 1 offers some example narratives.

Investigating narratives is important because, when combined with numerical ratings, their use increases the total amount of performance-related information being measured, and therefore, simultaneously increases the reliability and bandwidth of total measurement (Speer, 2018). Narrative comments and numerical ratings each reflect the construct of job performance and correlate meaningfully with one another, as raters often use the comments to justify rating decisions (Brutus, 2010; David, 2013). Thus, the inclusion of both measures increases the total amount of true score variance and, as such, is likely to increase reliability. Furthermore, although traditional numerical ratings and performance narratives each contain information about ratee behavior, narratives can offer additional insights not captured by numerical ratings, allowing raters to describe competencies, goal achievements, and the effects of an employee's behavior on others in rich and contextualized way (e.g., Brutus, 2010; Speer, 2018). Consistent with these points, when combined with traditional numerical ratings, narratives explain unique variance in employee promotions, turnover, and future performance ratings (Speer, 2018).

NLP has allowed researchers to automatically score performance narratives efficiently. Previous efforts have shown that NLP PA scores correlate with numerical performance ratings and future performance outcomes (Speer, 2018), can be used to identify what performance themes are discussed in text (Speer et al., 2019), and can identify gender differences in how managers write performance narratives (Doldor et al., 2019) and what challenges leaders face (Tonidandel et al., 2021). However, prior efforts have used older NLP methods (e.g., bag of words [BOW]) and have required substantial amounts of data to develop supervised NLP models.

<b>Example Performance Narrative Comments</b>
Overall, you had a great year, but there a few improvements that we can work on together to make this next year even better. You are great at following directions and working through them very specifically, but if a situation changes and you need to adapt your procedures without my help you often stumble. We've talked before about some fears that you have concerning being treated harshly after making decision at previous jobs, but I think I have shown you that that isn't the case here! I think if we are able to instill more confidence in you, then you will be able to rise above every occasion and perform stronger than ever.
Carol is the first to volunteer to help a coworker who is overburdened. She's always volunteering for extra assignments - because it helps her career, but also because she wants the company to succeed. She comes in every day with a can-do attitude and a good work ethic. I commend her for this.
Jenna has excellent knowledge of the practical application of the Labor Code and expects a certain knowledge of others in this respect too. Because the topic is not easy for everyone, Jenna repeatedly devotes herself to training and reminding others of key principles when working with others. Jenna carefully worked out the project of a potential change of supplier. It is necessary to obtain offers from several potential suppliers according to the specified parameters and evaluate the offers according to the given criteria. Jenna did a great job here and saved the company money by being detailed in supplier vetting. Overall, Jenna devoted herself to her goals and fulfilled them beyond the normal expectations. I expect more of the same in the year ahead.
This position does not demand that a person master all possible technical knowledge involved in the role, as that would be impossible. Yet, I do think we could work on upskilling for the next year. You have expertise in some of our products and a good basic knowledge of software needed for your role. More importantly, you have the ability to learn, as you demonstrate in the examples in your comments. However, I'd like to see you be more proactive in the future about learning new software on the horizon. Interpersonally, feedback from others indicates that when you work with someone, you develop a positive relationship with them. This is something I have noted you do consistently.
Justin is a " go to " person and is often reached out for advice and info. One of the leading resources of info and experience for all teams! Justin is willing to help out and comes in on short staffed situations with almost no or very little notice. Integrity is a priority for Justin, and he holds others no more accountable than he does for himself. Justin is proactive and not complacent as he assures things are carried out and takes the opportunity to correct things. The SDP initiative is one example but there have been many other examples. Additionally - customer satisfaction scores are 4.5 out of possible 5. Justin exceeds this, I have many examples listed in the perf notebook. Great job!

**Figure 1.** Example performance narrative comments.

## Assessing Performance Narratives Using Natural Language Processing

Prior to 2022–2023, most NLP applications within the organizational sciences used the BOW framework, where the simple presence of word phrases occurring in the text is the focus of the analysis. A typical BOW procedure (e.g., Speer, 2018, 2020) involves cleaning text (e.g., removing stopwords, standardizing punctuation) and then forming a document term matrix that represents the frequency of word phrases occurring in the text. Once operationalized as word vectors within this document term matrix, the vectors can be used as predictor input features into supervised ML algorithms (BOW-ML) to predict target scores (e.g., subject matter expert ratings [SME]).

More recently, neural network transformer models (e.g., Devlin et al., 2018; Liu et al., 2019; Vaswani et al., 2017) emerged as the dominant NLP architecture (Min et al., 2021). Transformers better encapsulate the contextual meaning of language, having been trained to predict words based on their context. Via specialized attention algorithms and deep neural networks, transformers better utilize text information. These dense neural networks can be used for language generation. Additionally, if the goal is to predict some target variable, such as interview ratings or performance evaluations, an upper neural network layer can be stacked above the language layers (i.e., language layers capture the meaning of language and allow for prediction and generation of language). This additional layer translates text representations into predicted target variable scores and, in this case, can be used to predict numerical PA ratings.

To our knowledge, transformers have not been applied to PA narratives in the organizational sciences literature, though they have outperformed BOW for other organizational tasks (e.g., Speer

et al., 2023; Thompson et al., 2023). In this study, we developed both supervised BOW and supervised transformer-based models to score performance narratives. Building on this, we then explored the efficacy of two off-the-shelf (i.e., zero-shot) LLMs: GPT-3.5 and GPT-4. Whereas previous transformer models (e.g., BERT from Vaswani et al., 2017; DeBERTa from He et al., 2020) typically require supervised fine-tuning for effective use in the organizational sciences, LLMs like GPT require no such fine-tuning. Instead, users can simply indicate the desired task they wish to perform. An example task is shown in Figure 2, where the user instructs the model to evaluate PA comments and score those comments on a 1–5 scale. This task is unsupervised because the model is not trained—meaning the model parameters are not updated—to recreate the target criterion. Rather, a “zero-shot” language-based prompt helps inform the model of the desired text analysis.

## Study Research Questions

For all three types of algorithms (supervised BOW, supervised transformer, off-the-shelf LLM), we expected strong levels of convergence with performance-related variables. In this study, we used two convergent ground truth measures: traditional numerical ratings and SME ratings of the text. A benefit of traditional numerical PA ratings is that they are often readily available in human resources systems, therefore resulting in large sample sizes when training the model. On the other hand, traditional numerical ratings serve different purposes than narrative comments (Brutus, 2010; Speer, 2018), and it may be worthwhile to train algorithms to directly evaluate the sentiment of written comments. Thus, we also used SME ratings of the text. The narrative itself contains cues that allow for inference of employee performance, and these scores serve as a more direct measure of an employee’s performance *based on the narrative text*. Strong correlations with these measures are required to establish evidence of construct validity.

*Research Question 1:* What is the correlation between NLP scores with (a) traditional numerical ratings and (b) SME ratings of narrative valence.

Supervised NLP methods are trained specifically to recreate the target ground truth scores. Thus, they are highly customized. In comparison, we use zero-shot LLM prompting to score the narratives “off-the-shelf.” Despite the lack of customization, LLMs have demonstrated remarkable capabilities

*“You will read a performance appraisal review that describes employee job performance over the course of a year. You are an attentive reader and expert performance appraisal assessor who wants to evaluate how positive or negative an employee’s performance is at work.*

*Read the performance appraisal review and categorize/rate the employee’s job performance using a 1-5 scale where*

*1 = poor performance (person needs to drastically improve) and*

*5 = excellent performance (is a star performer).*

*Category rating values are 1 = poor performance, 2 = performance needs work, 3 = average performance (i.e., meets expectations), 4 = performance is above expectations, 5 = performance is excellent.*

*Do not return text. Do not repeat the prompt. Only return an integer indicating your rating.”*

**Figure 2.** Prompt instructions used for GPT performance appraisal assessment.

of understanding and generating language and may perform well at evaluating performance valence despite not being explicitly trained to do so.

*Research Question 2:* How do psychometric properties (correlations with traditional numerical ratings, correlations with SME ratings) differ between supervised BOW, supervised transformer models, and off-the-shelf LLM scores?

## Methods

### Datasets

Six PA samples were included, coming from varied organizations and using diverse PA formats, therefore representing a large and assorted set of data for this study. Table 1 provides details regarding each sample. For samples where there were multiple comments nested within ratee (Samples 1 and 2), we concatenated all narrative comments into one overall comment per ratee. Because sample sizes varied across the six samples, and to avoid any one sample from dominating model training, sample size within subsample was restricted to 2000 by randomly selecting 2000 cases from that sample. Sample sizes for Samples 1–6 were 841, 110, 133, 2000, 366, and 1753. All respondents possessed an open-ended performance comment as well as a traditional numerical performance rating. As discussed below, a subset of narrative comments were also rated by SMEs, such that a random subset of 200 ratee comments per sample were randomly selected, or in the case where a sample  $N$  was less than 200, the total number of ratings. Sample sizes for Samples 1–6 for SME ratings are 200, 110, 133, 200, 200, and 200.

Supervised models were trained using k-folds cross-validation across all samples to avoid overfitting and to create generalizable algorithms. Our focus for this research was to develop algorithms to score overall job performance.

### Target Criterion Scores

Two sets of target criterion scores were examined, with these used to train supervised models and also serving as convergent measures of job performance. The first criterion was continuous, numerical performance ratings. The benefit of numerical ratings is that ratings naturally existed for all narratives, resulting in large sample sizes for model training and evaluation. Second, we used direct SME ratings of performance narrative sentiment. Due to the time-consuming nature of SME judgments, we reviewed a random subsample of 1,043 SME ratings across the six samples, sampling 200 ratees from each sample, or in the case where a sample  $N$  was less than 200, the total number of ratings (resulting in sample sizes of 200, 110, 133, 200, 200, and 200 for samples 1–6). The three study authors independently reviewed the performance narratives, with two raters randomly assigned to each comment. Ratings were made using a scale ranging from 1 to 5. Inter-rater agreement (G [q,1], Putka et al., 2008) averaged .88 across the six samples for a composite of two raters, and it averaged .79 for a single rater. These values are presented in Table 2. Table 2 also reports correlations between SME ratings and traditional numerical ratings within each sample ( $\bar{r} = .60$ ).

### Natural Language Processing Algorithms

*Supervised Bag of Words Scoring.* We trained a supervised BOW-ML algorithm to recreate traditional numerical performance ratings, with this BOW model serving as a comparative baseline. Because all ratee comments had both narrative and traditional numerical performance data, the total sample size was the sum of all the sample sizes across the six samples ( $N = 5203$ ). Data were trained using five-fold cross-validation randomly sampled across all six samples. This was used to

**Table 1.** Sample Descriptions.

Sample	Description
Sample 1	These data were the first sample described in Speer (2020). Managers were recruited from Amazon Mechanical Turk (MTurk) and made performance evaluations for two of their direct reports (their employee with letter of first name closest to letter “A” and then “Z”). Traditional numerical ratings and narratives were fully provided for 841 rated employees and for each of the Great 8 performance dimensions (Kurz & Bartram, 2002), as well as for overall job performance. Comments were aggregated within rater by concatenating their responses, and all of the traditional numerical performance ratings were averaged into a composite score to represent total job performance at the ratee level (for details see Speer, 2020).
Sample 2	This sample was the second described in Speer (2020). This sample used the same rating materials as sample 1. However, the ratings were upward, such that 110 raters assessed the job performance of their immediate supervisor. Raters were employed undergraduates at a Midwestern United States university. Once again, numerous comments were provided by each rater (eight for the Great 8 and one for overall job performance) and these were aggregated into a single comment for each ratee.
Sample 3	This sample was the third described in Speer (2020). Like Sample 2, raters were employed undergraduates at a large Midwestern United States university who made upward ratings of their immediate supervisors. Unlike Sample 2, the format of narratives was only a single open-ended text box. A total of 135 ratings were made.
Sample 4	This sample originates from Speer (2018). Ratings came from a large United States financial services organization. The PA evaluations were part of the annual performance review and were provided by immediate supervisors. The ratings were used for both administrative and developmental purposes. Over 15,000 performance narratives existed, but to minimize any sample from dominating the training dataset, a random subset of 2000 independent ratee comments were chosen for this research. A single overall job performance rating was provided for each ratee, which was accompanied by a free-form comment box (for details see Speer, 2018).
Sample 5	Sample 5 raters were currently employed managers who participated via Amazon Mechanical Turk. Participants only had access to the study if they indicated their job function was “management.” Upon entering the survey, respondents had to answer “yes” to currently being a direct supervisor of employees (i.e., manager) and having had completed a PA in the past year. Only respondents who met these criteria could continue the survey. Within the survey, managers rated three of their direct reports. Respondents were first provided a definition of PAs, and then they made ratings of their first direct report, who was chosen by identifying the employee with letter of first name closest to “A.” Upon completing ratings for that individual, managers rated employees closest to the letter “M” and then “Z.” Participants were told to make ratings as if they were part of the annual review, such that ratings were likely to influence pay, promotions, and terminations. Across all rated employees, there were a total of 366 narratives completed. A single narrative prompt (i.e., overall comment box) was provided for each ratee, and we analyzed overall job performance based on the composite of ratings across the Great 8 performance dimensions, similar to studies 1–3.
Sample 6	Annual performance ratings from a large multinational European wholesale company were used for Sample 6. Employees from all job levels were evaluated using an overall task performance narrative and numerical ratings of task effectiveness. There were 1753 performance ratings (i.e., ratees) analyzed in this study.

train the initial model and to compute scores for each holdout fold. Within each sample, performance ratings were standardized (i.e., z-scored) prior to training.

Pre-processing was done using the procedures outlined by Speer (2020), which involved typical pre-processing procedures such as lowercasing, removing select punctuation, replacing contractions,

**Table 2.** Properties of SME Ratings and Traditional Numerical Ratings.

Sample	SME reliability G (q,l)	r SME w/ Num performance ratings
Sample 1	.81	.76
Sample 2	.80	.77
Sample 3	.85	.75
Sample 4	.73	.54
Sample 5	.82	.36
Sample 6	.74	.43
Average	.79	.60

Note. G (q,l) represents single rater reliability (Putka et al., 2008). Sample sizes for Samples 1–6 for traditional numerical ratings are 841, 110, 133, 2000, 366, and 1753. Sample sizes for Samples 1–6 for SME ratings are 200, 110, 133, 200, 200, and 200. SME = subject matter expert.

controlling for negation (e.g., “not great” becomes “no\_great”), removing a custom set of stopwords, and lemmatizing (Speer, 2020). After pre-processing, we ran ridge regression (e.g., Hoerl & Kennard, 1970; Zhou & Hastie, 2005) to recreate numerical performance ratings.

*Supervised Transformer Models.* Two supervised transformer models were trained. The first (initial transformer) was trained as a comparison to the BOW model, such that like the BOW-ML model, the initial transformer model was trained to recreate traditional numerical performance ratings. Identical cross-validation and fold settings were used.

We computed supervised transformer models using Python and Google Colab via the HuggingFace interface (Wolf et al., 2019). DeBERTa was used as the model architecture (He et al., 2020), which is based on the common BERT transformer architecture (Vaswani et al., 2017) but with improved transformer attention mechanisms that lead to improved model performance. The DeBERTa parameters have already been pre-trained on massive datasets, thus mitigating the need for large primary samples. Instead, transfer learning was applied in this research (e.g., Howard & Ruder, 2018; Wolf et al., 2019) by minimally tweaking the weights on the newly collected data from this study. To prevent drastic model forgetting when updating parameters, we fixed the embedding layer and the bottom eight language layers, meaning we held those parameters constant and only allowed for parameter updates to the remaining upper layers. To make model updates, we used established model hyper-parameters from a similar transformer project we had worked on as starting values (weight decay = .0, learning rate = .00005, four epochs, batch size = 8). These are very similar to the DeBERTa defaults, and some additional exploration of changes to these hyper-parameters did not lead to any notable model improvements. For example, changes to weight decay and learning rate did not result in improvements in training loss, and four epochs were found to be ideal. These settings were applied to train the initial transformer model to recreate numerical performance ratings.

Second, we trained a final and updated model customized to predict SME judgments of narrative performance valence, once again using k-folds cross-validation. Rather than training a new model entirely, we used the initial transformer model parameters as the starting values and then gradually tweaked them through four additional passes (i.e., epochs) through the data (note the choice to use the starting values from the initial transformer model was inconsequential, as follow-up analyses revealed that this choice only improved model psychometrics minimally, such that correlations between NLP scores and SME ratings only differed .01 on average). Once completed, the finalized transformer model was thus (a) pretrained using the default DeBERTa parameters, (b) fine-tuned to predict traditional numerical performance ratings, and (c) further fine-tuned to predict SME ratings of the performance narratives.



To assist researchers and practitioners who wish to perform similar NLP work themselves, we have included code for how to perform supervised DeBERTA, contained in the following OSF link: [https://osf.io/nzva8/?view\\_only=f969d75c21484aaca47bba51db796d6](https://osf.io/nzva8/?view_only=f969d75c21484aaca47bba51db796d6). Additionally, we have included a mock PA dataset so users can experiment themselves.

*Off-the-Shelf LLM.* We used the OpenAI API to generate NLP scores using both GPT-3.5 and GPT-4. Analyses were run in March 2024 using the prompt presented in Figure 2. The prompt results in a single NLP score for each narrative text. Like above, we have also provided OSF code so that researchers and practitioners can perform similar work themselves.

### Validation Strategy

The supervised NLP scores were all independent of model training given the nested k-folds cross-validation. Within each sample and for all methods, NLP scores were correlated with traditional numerical ratings and SME ratings of the same comments.

## Results

A summary of results across the six samples can be found in Figure 3, and Table 3 provides more detailed results by sample.

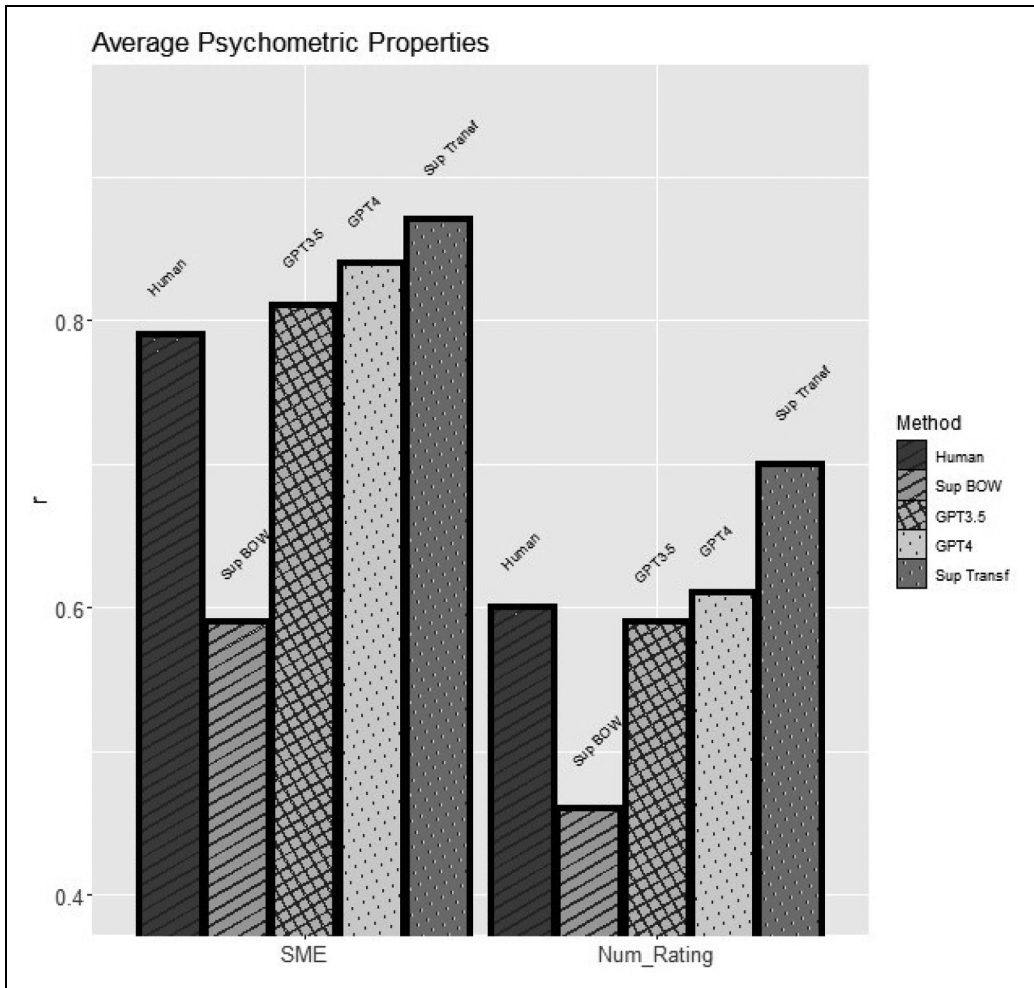
### Supervised Bag of Words Versus Supervised Transformer

The BOW model and initial transformer model were each trained to recreate numerical performance ratings and, therefore, serve as direct comparisons. Initial transformer scores had higher correlations

**Table 3.** Correlations by Model.

Sample	Bag of words model	Starting transformer model	Final transformer model	GPT3.5	GPT4
<i>r</i> SME ratings					
Sample 1	.67	.85	.92	.87	.88
Sample 2	.52	.87	.91	.87	.84
Sample 3	.55	.86	.91	.88	.90
Sample 4	.66	.73	.85	.77	.81
Sample 5	.64	.81	.90	.86	.90
Sample 6	.48	.61	.76	.60	.70
Average	.59	.79	.87	.81	.84
<i>r</i> Num performance ratings					
Sample 1	.61	.81	.82	.77	.78
Sample 2	.62	.76	.84	.80	.79
Sample 3	.39	.74	.77	.72	.76
Sample 4	.53	.64	.69	.49	.49
Sample 5	.28	.45	.47	.41	.42
Sample 6	.35	.48	.59	.38	.44
Average	.46	.65	.70	.59	.61

*Note.* Shown are holdout correlations independent of model training. BOW = bag of words. SME = subject matter expert. The Starting Transformer Model was trained solely to predict traditional numerical performance ratings. The final transformer model was pre-trained to predict traditional numerical performance ratings and then trained for several additional epochs to predict human subject-matter-expert ratings. Sample sizes for Samples 1–6 for traditional numerical ratings are 841, 110, 133, 2000, 366, and 1753. Sample sizes for Samples 1–6 for SME ratings are 200, 110, 133, 200, 200, and 200. Single-rater inter-rater agreement of SME ratings for Samples 1–6 are .81, .80, .85, .73, .82, and .74, for an average of .79.



**Figure 3.** Average psychometric properties by scoring method.

Note. Results are averaged across the six samples.  $r$  = correlation; SME = subject matter expert rating of the text; Num\_Rating = traditional numerical rating; Human = human SME values. In respect to the SME result it is the average single-rater reliability, and in respect to the Num\_Rating results it is the average correlation with traditional numerical ratings. Sup BOW = supervised BOW; Sup Transf = supervised transformer.

with SME ratings ( $\bar{r}$  .59 for BOW,  $\bar{r}$  .79 for initial transformer model, Steiger  $z = 11.69, p < .01$ ) and higher correlations with traditional numerical ratings ( $\bar{r}$  .46 for BOW,  $\bar{r}$  .65 for the initial transformer model, Steiger  $z = 20.92, p < .01$ ). Thus, by using a more sophisticated NLP algorithm, psychometric properties were improved, providing support for transformer-based scoring in this context.

### Comparing Supervised Transformers

The initial transformer model was trained to recreate numerical performance ratings, whereas the final transformer model was trained to assess the valence of performance narrative comments by further fine-tuning the model using SME ratings. Scores from the final transformer model correlated on average .87 with SME ratings, versus .79 for the starting transformer model (Steiger  $z = 9.79, p < .01$ ). It should be noted that the value of .87 is larger than the reliability for a *single* human rater (.79).

Interestingly too, convergence with traditional numerical ratings was improved for the final transformer model (.70 vs. .65). Given superior performance overall, we focus on the finalized supervised transformer going forward. The model is made freely available: [https://osf.io/nzva8/?view\\_only=f969d75c21484aaea47bbea51db796d6](https://osf.io/nzva8/?view_only=f969d75c21484aaea47bbea51db796d6).

Before moving on, it is worth noting how much larger the correlations were between NLP scores and SME ratings than NLP scores with traditional numerical ratings. However, this is a common occurrence for ML solutions, as well as for multi-construct multi-method contexts (e.g., Hoffman et al., 2010; Lance et al., 2000; Lievens et al., 2006), such that ML scores exhibit larger correlations when the data that inform the ML and target scores come from the same source (e.g., Hickman et al., 2022; Koutsoumpis et al., 2024; Perrotta et al., 2023). For example, Hickman et al. (2022) developed separate ML models to recreate SME ratings of interview responses, as well as models to predict self-report personality ratings from the interview responses. The average correlation for the former scenario was .40, whereas it was just .12 in the latter.

In the case of SME ratings, both the ML scores and the SME ratings are directly informed from the narrative data (i.e., similar method). On the other hand, in cases where ML is used to predict self-report data, the data used to inform the ML score (narratives) is different than the data used to inform the target scores (i.e., different method), and as such these measures are likely to reflect slightly different factors. This is consistent with past multi-construct multi-method research, such that scores from different sources of data often exhibit lower convergence (e.g., Hoffman et al., 2010; Lance et al., 2000; Lievens et al., 2006). In the case of PAs specifically, the traditional numerical ratings are not based directly on the narratives and serve a different functional purpose in practice (Speer, 2018). Reflecting this, the correlation between these data sources was just .60 in the current study. Thus, the traditional numerical ratings and SME scores were related, but clearly reflected different variance. The traditional numerical ratings serve only as indirect measures for ground truth and are used for different purposes than the narrative themselves (Brutus, 2010; Speer, 2018), whereas SME judgments more directly represent the valence of narrative comments.

An additional reason why correlations with target scores may have been higher for SME ratings than for traditional numerical ratings is that the reliability of SME target scores was higher than that for traditional numerical ratings. The SME composite score had an average reliability of .88, whereas the best estimate of single-rater supervisor reliability is .65 (Speer et al., 2024). However, despite reliability differences, the evidence from this study, as well as research in other contexts, does not support this as the sole reason for differential correlations with target scores. If this study's final transformer model correlations are corrected for unreliability, the correlations are .87 for traditional numerical ratings and .93 for SME ratings,<sup>1</sup> thus still demonstrating a difference. Data from other studies are consistent with this as well. For example, if the data from Hickman et al. (2022) are corrected for unreliability, correlations with target scores are .49 for SME ratings and .13 with self-report composite scores. Similarly, corrected correlations from Koutsoumpis et al. (2024) are .61 for SME target scores and .36 for self-report composite scores. Taken together, although target score reliability is one potential explanation for differential correlations with target scores, in general, correlations are likely to be larger when the data that inform the ML and target scores come from the same source.

That said, it should be noted that the degree of convergence between NLP scores and SME ratings coincided with SME reliability. Four samples had single rater reliability .80 or above, and in these samples, the NLP correlation with SME ratings was greater than .90. On the other hand, the lowest two correlations between NLP scores and SME ratings (.85 in Sample 4, .76 in Sample 6) coincided with the two samples where SME reliability was lowest (.73 and .74). This makes conceptual sense, as NLP scores cannot recreate scores that are themselves unreliable, and thus to an extent, the NLP scores perform better at things that human raters are themselves good at. In respect to Samples 4 and 6, where the weaker reliability and weaker convergence occurred, the PA comments were often quite complex and there was variability in how they were written. Managers varied greatly

in their use of company-specific jargon, whether the comment was used to list objective metric performance, whether the comment was used for developmental tips, and the general tone of the comment. The PA data from Samples 4 and 6 were also used for administrative decisions, which increases the likelihood that bias and error affect PA data (e.g., Jawahar & Williams, 1997; Speer et al., 2020). Thus, the NLP scores exhibited weaker psychometric properties in samples where they would be expected to do so.

### *Off-the-Shelf LLM Versus Supervised Transformer*

Table 3 contains results for GPT-3.5 and GPT-4 scores. GPT-3.5 had superior performance when compared to BOW, correlating .81 with SME ratings (Steiger  $z = 14.12, p < .01$ ) and .59 with traditional numerical ratings (Steiger  $z = 11.88, p < .01$ ). These values were also similar, though lower, to those seen with the final transformer model, with the difference in correlations with SME ratings being .06 (Steiger  $z = 7.31, p < .01$ ) and the difference in correlations for traditional numerical ratings being larger ( $r_{diff} = .11$ , Steiger  $z = 19.83, p < .01$ ). Still, it is noteworthy that GPT-3.5 scores had higher reliability than single rater SME judgments (.79) and approached psychometric performance of the fully supervised DeBERTA scores, given GPT-3.5 requires zero training to score the narratives and can be used free of charge via the ChatGPT interface.

Similarly, GPT-4 performed well and was a slight improvement over GPT-3.5. The average correlation with SME ratings was .84 (Steiger  $z = 3.52, p < .01$ ), and it was .61 for traditional numerical ratings (Steiger  $z = 3.09, p < .01$ ). These values are similar to the psychometric properties of the supervised transformer model, though slightly lower. GPT-4.0 reliability was also higher than single rater human evaluations (.79), and the scores exhibited reasonable psychometric properties for all six samples, thus demonstrating evidence of generalizability across PA prompts, jobs, and organizations. Lastly, the average correlation between GPT-4 scores and supervised transformer scores was .86, demonstrating convergent validity evidence (e.g., Furr, 2021).

## **Discussion**

Supervised NLP has commonly been used to recover a priori-targeted constructs from text in the organizational sciences. Although supervised NLP generally exhibits favorable psychometric properties, models are resource-intensive to develop. Newer off-the-shelf LLMs present an opportunity to assess organizational constructs without the need for developing customized models, and yet the validity of such approaches is unknown. The current research tested whether off-the-shelf LLMs accurately assess job performance from narrative comments, pitting them against customized, supervised transformer models and testing the generalizability of LLM scores across six PA datasets and thousands of PA comments. There are several major findings and contributions from this work. We present those here, and we highlight a decision process for considering how to use off-the-shelf LLMs in practice.

The supervised transformer model demonstrated impressive psychometric evidence, with an average correlation of .87 with SME ratings and .70 with numerical ratings made during the appraisal. The fact that the correlation with SME ratings was .87 and the reliability for a single human coder was only .79, speaks to the potential benefits of using NLP to score performance narratives. Not only does NLP eliminate the need for time-consuming and expensive human coding, but NLP algorithms also do not suffer from some of the challenges associated with human judgments. Humans are susceptible to distraction, influenced by changes in mood, inconsistent in the rules used to form judgments, and are likely to only consider portions of a comment when making evaluations. On the other hand, an NLP algorithm is consistent in how it derives scores, considers all available information, and does not fatigue. As such, NLP can achieve not only cost savings but

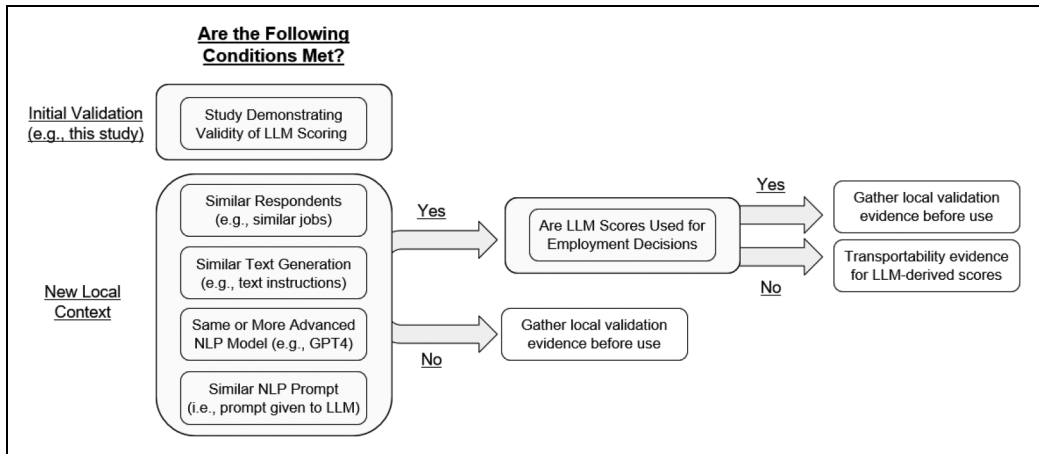
also psychometrically sound measurement. For these reasons, we make our supervised transformer algorithm freely available to the research community.

Yet, with the rapidly widespread adoption of off-the-shelf LLMs, customized supervised NLP models may have less use going forward. We found there to be similar though slightly worse performance between customized supervised transformer models and the output from GPT-3.5 and GPT-4, with GPT-4 producing slightly more accurate scores than GPT-3.5. The average GPT-4 correlation was .84 with SME ratings and .61 with traditional numerical ratings, with each similar, though slightly lower, than the supervised transformer. Taken together, we conclude that GPT-4 provides similarly accurate PA scores as supervised transformer models across these diverse PA contexts, though if users are concerned with achieving the highest levels of psychometric performance, using supervised transformers will likely yield slight improvements.

In respect to the larger consideration of using LLMs within the organizational sciences, we believe these results support off-the-shelf LLM scoring within PAs specifically, with a caveat. Namely, if the context requires strong evidence of validity, evidence from this study should only be used to support off-the-shelf applications in contexts that are similar to the current validation study, consistent with transportability validity evidence (Principles for the Validation and Use of Personnel Selection Procedures, 2018). Issues of transportability have generally been applied to hiring contexts, but the logic extends to the validation of any set of scores, including across varied organizational sciences contexts (both applied settings and research contexts). Based on a transportability argument, if trying to implement an assessment within a new job or context, then absent local validation evidence (e.g., that the assessment scores are related to important criteria; a measure exhibits strong convergent correlations with target scores), the assessment lacks validation evidence. However, if the assessment has shown to be valid elsewhere (e.g., in other jobs, contexts), then in such situations, an organization or researcher may rely upon transportability evidence to justify the assessment based on results of a previous validation research study and assuming similarity between the two contexts of use. We introduce Figure 4 as a decision chart for how users might adapt the results of this study to assess PA narratives and other organizational constructs of interest with off-the-shelf LLMs, offering advice for those interested in using LLMs for employment decisions (e.g., hiring, promotion decisions) or for lower-stakes research purposes. Regarding this last point, a lower standard for application occurs when LLMs are used to score text in low-stakes contexts such as for research. In such settings, confidence in measurement validity is still important, but there are no legal standards that must be met in the same way that are required for high-stakes assessment usage.

As outlined in Figure 4, to confidently use off-the-shelf LLMs to assess psychological or organizational constructs absent local validation evidence, it is recommended that (1) a validation study (i.e., study containing evidence that the LLM-generated scores are valid) has been conducted elsewhere demonstrating that off-the-shelf LLM scores display acceptable psychometric properties (e.g., approximate the human SME correlation with target scores). (2) If this is satisfied, the new context in which off-the-shelf LLMs will be used should be similar to the original validation context, with (a) similar respondents (e.g., applied to similar jobs, applied to incumbents, all text provided by trained respondents) and (b) similar text generation instructions (e.g., the PA survey instructions for text are similar, same motivation for surveys). Likewise, (c) the same NLP model and (d) NLP prompt should ideally be leveraged. Under these conditions, it is reasonable to assume the contexts are similar enough such that an off-the-shelf LLM shown to produce valid scores in a prior context will also produce valid scores in the new context. Absent these conditions being met, there is not strong evidence that off-the-shelf LLM scores will accurately measure the intended construct, and as such, off-the-shelf LLM scores should first be validated in that context (e.g., correlated with SME ratings) prior to implementing the LLM for use (either research or operational purposes).

As an example, let us assume we validated off-the-shelf LLM scores on data from study 1 with the previously used GPT-4 model and prompts. However, our new context is Sample 6, which differs in



**Figure 4.** Transportability decision chart for using large language models to score psychological and organizational sciences constructs.

Note. LLM = large language model. New Local Context refers to new environment where user wishes to generate LLM scores. Similarity in New Local Context is not a strict dichotomy, and users should evaluate context similarity requirements based on intended use (e.g., high-stakes vs. low-stakes).

the text generation instructions (from numerous narrow prompts to a single broad prompt). Let us also assume we used GPT-3.5 due to cost and leveraged a different prompt (“Score this performance appraisal narrative on a scale from 1 to 5, where 5 = great”). Under the original validation context, NLP correlations with SME ratings and traditional numerical ratings were .88 and .78, thus exhibiting strong psychometric properties. However, under the new scenario, the values are just .58 and .36. Thus, results from the original validation did not “transport” to the new context (i.e., did not generalize to), and this is because the contexts are not similar enough to infer transportability. This does not necessarily mean that off-the-shelf LLM scores will not be valid in new contexts absent transportability evidence. However, it does mean that users cannot be as confident that the derived scores will, in fact, be valid.

In respect to these points, it is worth discussing just how similar NLP prompts must be to retain transportability evidence. NLP prompts can affect NLP output, and it is not always clear what will and what will not work (Meincke et al., 2024). There are likely certain prompt features that should remain stable to maintain confidence in transportability, particularly if relying on transportability evidence for high-stakes employee decisions. For evidence of transportability, the core instructions should not greatly change. For example, in our study, we instructed GPT to rate performance on a 1–5 scale. If we had changed this to a 1–3 scale, we would have found slightly different psychometric properties (i.e., convergence with numerical ratings was 16% lower and convergence with SME ratings was 12% lower).

On the other hand, other prompt features might be altered with minimal repercussions. These include trivial wording adjustments that don’t affect the underlying meaning, such as preferring “performance review” over “performance appraisal,” or minor specification changes, like requesting a numerical output. Additionally, omitting stylistic elements, like directing against returning text or echoing the prompt for cleaner output, likely have minimal impacts on transportability. However, it’s worth noting that these stylistic choices might have benefits in terms of output clarity and prevent the model from narrowing its focus prematurely, potentially affecting the final output’s relevance.

It should also be noted that LLMs change rapidly, and tomorrow's models will be more powerful than today's models. Figure 4 suggests that the same model should be used to establish transportability of LLM scores. However, if a more advanced model is created and then later used, it is likely that this model will have higher validity, and therefore the requirement to use the same LLM may not always hold. This said, different LLMs often will perform better at different tasks, and therefore without firm empirical evidence having been established, it will be unclear how valid LLM-derived scores are in such cases. For low stakes applications of LLMs (e.g., research), it may be reasonable to assume that a more advanced LLM will remain valid when compared to validity evidence collected elsewhere using a less advanced LLM.

On the other hand, for high stakes decisions, practicing caution and requiring the collection of local validation evidence in such settings is prudent. To this point, we add a final caveat regarding off-the-shelf LLM usage particularly for employment decisions (farther right on Figure 4). If used for employment decisions, off-the-shelf LLM scores would then be subject to laws and regulations such as the Civil Rights Act of 1964. If adverse impact were found, score validity would then be held to higher levels of scrutiny. We are not implying that a transportability approach would definitively fail in such a scenario, but given the higher stakes of assessment use, local validation is recommended in such a context. Thus, whenever off-the-shelf LLM scores are used for employment decisions, or in cases where sound transportability data does not exist, we recommend conducting local validation, at least until firmer case law is established.

Figure 4 can also guide usage in lower stakes research contexts, for which LLMs can be used to assess a whole range of constructs (e.g., work attitudes, employee emotions, leadership styles). Concepts of validity are vital to supporting research conclusions, and as such we recommend that users strive to achieve the aforementioned transportability requirements. However, slight violations might be tolerated in some research contexts, with the degree of acceptable violation depending on the intended use of LLM scores. For example, let's say LLMs were highly related to SME ratings of positive affect derived from employee emails. Another researcher may wish to assess positive affect in a new research context that also uses emails. The sample is very different from the original validation context though, using higher-level managers instead of lower-level employees. This decreases confidence in transportability. Yet, by using the same type of LLM, the same medium (i.e., email), the same NLP prompts, and given it is not a litigious context, there are likely enough similarities to make arguments of transportability and conclude that the LLM scores are reasonably valid in the new context.

### *Limitations and Areas for Future Research*

A positive feature of this study was that we were able to examine performance narratives across different PA contexts and organizations. Nonetheless, organizations differ in PA features such as company culture, training, initiatives to facilitate the PA process, pay allocation norms, and so forth, which all may impact how narratives are written and, therefore, how algorithms perform. As such, it is possible our findings will not generalize to all PA contexts. If using a supervised model, this is of less concern. Practitioners could even take the algorithm we shared and further fine-tune it on their local data.<sup>2</sup> This option is likely to yield strong validity. Yet, this defeats some of off-the-shelf LLM benefits discussed in this paper. Ultimately, we believe Figure 4 addresses these concerns when considering future off-the-shelf LLM use to assess psychological and organizational constructs.

In relation, there is an intermediate between off-the-shelf LLMs and supervised modeling—that is, to fine-tune the LLM on local text data prior to prompting it. GPT is trained across a massive and broad set of text, but if the local intended context for use differs greatly in the type of language used (e.g., interview responses), it can be beneficial to update model parameters based on context-specific language (e.g., Demszky et al., 2023; Wu et al., 2023). Because this contradicts this

study's goals of understanding off-the-shelf performance of LLMs, it was not performed. However, this is a promising area for future research.

Finally, this study only examined PA data. There are many other sources of unstructured organizational text, the likes of which vary in complexity and purpose. It is unclear if our pattern of findings will generalize to other organizational use cases. Thus, future research is needed, and particularly to flesh out whether the proposed transportability implications function satisfactorily across varied organizational contexts.

## Conclusion

We found that off-the-shelf LLMs demonstrated similar, though slightly lower, psychometric properties as supervised transformer models in assessing PA narratives. Given the similarity in psychometric performance to supervised transformers and given that LLM scores were more reliable than single rater SMEs, this suggests that off-the-shelf LLMs appear to be psychometrically sound methods to evaluate PA text. We presented a framework for interpreting psychometric evidence when using off-the-shelf LLMs within other contexts and when assessing other organizational constructs. Taken together, we believe this research establishes off-the-shelf LLMs as a useful method for researchers and practitioners, and we encourage users to apply and test our proposed decision rules for applying off-the-shelf LLM to score organizational constructs.


## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Andrew B. Speer  <https://orcid.org/0000-0002-3376-2103>

## Notes

1. A reviewer inquired about the magnitude of the corrected correlations found here, given they approached 1.0. These values are quite high and stand in contrast to corrected correlations commonly seen in other domains of the organizational sciences where more traditional scoring is used, such as personnel selection (e.g., Sackett et al., 2022). The major difference for contexts such as personnel selection is that there is not a 1–1 construct match between the predictor content (e.g., test content assessing knowledge, skills, abilities, and other characteristics, KSAOs) and the target domain (often job performance). Job performance, which is commonly the dependent variable in personnel selection, cannot be closely predicted by KSAO data alone, as job performance is influenced by a complex blend of factors such as KSAOs, but also supervisor leadership abilities, motivation, employee training, job design, job attitudes, compensation, among many others (cf. LeBreton et al., 2014). On the other hand, a performance narrative is explicitly designed to describe the construct of job performance, and in many cases, narratives describe performance in straightforward and easily scorable ways. Because of this, higher levels of convergence can be found.
2. Researchers could collect target scores, such as a few hundred SME ratings of the text, and then gradually tune the model to update parameters on the local data. Such a process balances the power of large samples, previously fine-tuned algorithms, and then further tuning to maximally customize an algorithm for local organizational use.



## References

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., & Sutskever, I. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review*, 20(2), 144-157. <https://doi.org/10.1016/j.hrmr.2009.06.003>
- Campion, M. A., & E. D. Campion. (2023). *Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research*. Personnel Psychology.
- David, E. M. (2013). Examining the role of narrative performance appraisal comments on performance. *Human Performance*, 26(5), 430-450. <https://doi.org/10.1080/08959285.2013.836197>
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2, 688-701. <https://doi.org/10.1038/s44159-023-00241-5>
- DeNisi, A. S., Murphy, K. R., Lee, K., & Toutanova, K. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology*, 102(3), 421-433. <https://doi.org/10.1037/apl0000085>.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Doldor, E., Wyatt, M., & Silvester, J. (2019). Statesmen or cheerleaders? Using topic modeling to examine gendered messages in narrative developmental feedback for leaders. *The Leadership Quarterly*, 30(5), 101308. <https://doi.org/10.1016/j.leaqua.2019.101308>
- Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*, 108, 1277-1299. <https://doi.org/10.1037/apl0001082>
- Furr, R. M. (2021). *Psychometrics: An introduction*. Sage publications.
- Gorman, C. A., Meriac, J. P., Roch, S. G., Ray, J. L., & Gamble, J. S. (2017). An exploratory study of current performance management practices: Human resource executives' perspectives. *International Journal of Selection and Assessment*, 25(2), 193-202. <https://doi.org/10.1111/ijsa.12172>
- He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(8), 1323-1351. <https://doi.org/10.1037/apl0000695>
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Journal of Technometrics*, 12(1), 55-67. <https://doi.org/10.1080/00401706.1970.10488634>
- Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology*, 63(1), 119-151. <https://doi.org/10.1111/j.1744-6570.2009.01164.x>
- Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv:1801.06146*.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R* (7th ed.). Springer.
- Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, 50(4), 905-925. <https://doi.org/10.1111/j.1744-6570.1997.tb01487.x>
- Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., Speer, A., Hardy, J. H., Gibson, C., Frost, C., Liu, M., McNeney, D., Capman, J., & Lowery, S. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology*. Advance online publication. <https://doi.org/10.1111/peps.12608>

- Koutsoumpis, A., Ghassemi, S., Oostrom, J., Holtrop, D., Van Breda, W., & de Vries, R. E. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning. *Computers in Human Behavior, 154*, 1-18. <https://doi.org/10.1016/j.chb.2023.108128>
- Kurz, R., & Bartram, D. (2002). Competency and individual performance: Modeling the world of work. In I. T. Robertson, M. Callinan, & D. Bartram (Eds.), *Organizational effectiveness: The role of psychology* (pp. 227-255). Wiley.
- Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13*, 323-353. [https://doi.org/10.1207/S15327043HUP1304\\_1](https://doi.org/10.1207/S15327043HUP1304_1)
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72. <https://doi.org/10.1037/0033-2909.87.1.72>
- LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*(4), 478-500. <https://doi.org/10.1111/iops.12184>
- Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91*, 247-258. <https://doi.org/10.1037/0021-9010.91.2.247>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.
- Meinke, L., Mollick, E. R., & Terwiesch, C. (2024). Prompting diverse ideas: Increasing AI idea variance. *arXiv preprint arXiv:2402.01727*.
- Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.
- Perrotta, J., Delacruz, A. Y., Tenbrink, A. P., Wegmeyer, L. J., Chawota, T., Mwangale, Z., & Speer, A. B. (2023). Text scoring work attitudes and perceptions: A performance comparison of procedures [poster presentation]. In Society for Industrial and Organizational Psychology Annual Conference, Boston, MA, USA.
- Principles for the Validation and Use of Personnel Selection Procedures. (2018). *Industrial and organizational psychology. Perspectives on Science and Practice, 11*, 2-97. <https://doi.org/10.1017/iop.2018.195>
- Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93*(5), 959-981. <https://doi.org/10.1037/0021-9010.93.5.959>
- Rottman, C., Gardner, C., Liff, J., Mondragon, N., & Zuloaga, L. (2023). New strategies for addressing the diversity–validity dilemma with big data. *Journal of Applied Psychology, 8*(9), 1425-1444. <https://doi.org/10.1037/apl0001084>
- Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology, 107*(11), 2040-2068. <https://doi.org/10.1037/apl0000994>
- Sartori, G., & Orrù, G. (2023). Language models and psychological sciences. *Frontiers in Psychology, 14*. <https://doi.org/10.3389/fpsyg.2023.1279317>
- Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology, 71*(3), 299-333. <https://doi.org/10.1111/peps.12263>
- Speer, A. B. (2020). Scoring dimension-level job performance from narrative comments: Validity and generalizability when using natural language processing. *Organizational Research Methods, 24*(3), 572-594. <https://doi.org/10.1177/1094428120930815>
- Speer, A. B., Delacruz, A. Y., Wegmeyer, L. J., & Perrotta, J. (2024). Meta-analytical estimates of interrater reliability for direct supervisor performance ratings: Optimism under optimal measurement designs. *Journal of Applied Psychology, 109*(3), 456-467. <https://doi.org/10.1037/apl0001146>

- Speer, A. B., Perrotta, J., Tenbrink, A. P., Wegmeyer, L., Delacruz, A. Y., & Bowker, J. (2023). Turning words into numbers: Assessing work attitudes using natural language processing. *Journal of Applied Psychology, 108*(6), 1027-1045. <https://doi.org/10.1037/apl0001061>
- Speer, A. B., Schwendeman, M. G., Reich, C. C., Tenbrink, A. P., & Siver, S. R. (2019). Investigating the construct validity of performance comments: Creation of the great eight narrative dictionary. *Journal of Business and Psychology, 34*, 747-767. <https://doi.org/10.1007/s10869-018-9599-9>
- Speer, A. B., Tenbrink, A. P., & Schwendeman, M. G. (2020). Creation and validation of the performance appraisal motivation scale (PAMS). *Human Performance, 33*(2-3), 214-240. <https://doi.org/10.1080/08959285.2020.1776713>
- Thompson, I., Koenig, N., Mracek, D. L., & Tonidandel, S. (2023). Deep learning in employee selection: Evaluation of algorithms to automate the scoring of open-ended assessments. *Journal of Business & Psychology, 38*, 509-527. <https://doi.org/10.1007/s10869-023-09874-y>
- Tonidandel, S., Summerville, K. M., Gentry, W. A., & Young, S. F. (2021). Using structural topic modeling to gain insight into challenges faced by leaders. *The Leadership Quarterly, 33*(5), 1-20.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 5998-5600*. <https://doi.org/10.48550/arXiv.1706.03762>
- Wolf, T., L Debut., V Sanh., J Chaumond., D Clement., A Moi., P Cistac., T Rault., R Louf., M Funtowicz., J Davidson., S Shleifer., P von Platen., C Ma., Y Jernite., J Plu., C Xu., T La Scao., S Gugger., A. M Rush. (2019). HuggingFace's transformers: State-of-the-art natural language processing. *arXiv:1910.03771*.
- Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., & Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Zhou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological), 67*(2), 301-320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

## Author Biographies

**Andrew B. Speer** is an assistant professor in the Kelley School of Business at Indiana University. His research deals with employee selection, personality, performance management, employee turnover, artificial intelligence, natural language processing, and machine learning.

**James Perrotta** is a doctoral student of industrial-organizational psychology at Wayne State University in Detroit, Michigan, and a research scientist at DDI. His research deals with integrating natural language processing and machine learning with areas such as employee selection and assessment, employee turnover, and adverse impact.

**Tobias L. Kordsmeyer** is a postdoctoral researcher in the Institute of Psychology at the University of Goettingen. His research focuses on person perception, sexual selection, behavioural endocrinology, and personality effects in work contexts.