International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

# Leveraging large language models for automated performance appraisals: Opportunities and challenges

Sri Kuchibhotla *

*Independent Researcher, Columbus, Ohio*

International Journal of Science and Research Archive, 2025, 14(03), 1268-1273

Publication history: Received on 04 January 2025; revised on 18 March 2025; accepted on 20 March 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.3.0803

## Abstract

One major issue with traditional performance appraisals is inefficiency, bias and subjectivity. Oftentimes large language models (LLMs) like GPT-4 offer a promising approach to standardize performance evaluations which leverage structured and unstructured feedback for data-driven assessments. In this study, a data set with structured and unstructured data is taken and fed into GPT-4 to analyze self-evaluations and mid-year performance reviews to automate the appraisal process and compare it to human evaluations. Although GPT-4 is generally accurate and is similar to human assessment, the main challenge lies in the non-quantifiable factors such as workplace dynamics and lack of emotional intelligence. Although AI models have a much more accurate prediction rate than manual performance appraisals, there is always a need for a human-in-the-loop (HITL) approach to help AI perform better. This study focuses on how human-in-the-loop (HITL) can help AI-based performance appraisals by bringing in non-quantifiable factors such as workplace dynamics and conflict resolution within the employee data.

**Keywords:** Artificial Intelligence; Large Language Models in HR; Performance Appraisals; Human Resources; AI Bias; Human-in-the-Loop

## 1. Introduction

Performance appraisals are a fundamental component of human resource management, enabling organizations to assess employee contributions, provide feedback, and guide career development. Traditionally, these evaluations rely on managerial reviews, peer feedback, and self-assessments, which is time-consuming, subjective, and prone to biases. LLMs can process and analyze large volumes of structured and unstructured data, identify patterns in employee performance, and generate objective, data-driven appraisals. Recent advancements in the field of Artificial Intelligence (AI) has opened impressive opportunities to automate the tedious tasks such as performance appraisals to offer promising alternatives for the traditional approaches. Large Language Models such as GPT-4 is great at analyzing large volumes of data, particularly structured and unstructured data, to identify patterns in performance and generate data driven results

### 1.1. Problem Statement

Traditional performance assessment methods are usually criticized for bias, time-consuming, and error-prone (DeNisi & Murphy, 2017). Managers may miss key achievements or judge based on current performance, self-rating, or personal biases, which leads to mistakes and employee dissatisfaction (Giles et al., 2021). These flaws may undermine the effectiveness of appraisals and affect worker motivation and turnover (Aguinis, 2019).

Artificial intelligence (AI) and particularly, large language models (LLMs) development is a potential solution by streamlining and making performance evaluations standard and automated (Nyathani, 2023). AI-based performance

* Corresponding author: Sri Kuchibhotla

reviews can reduce human biases and can be more cost-effective, offering a more objective, data-oriented assessment of employee performance (Gupta & Tembhurnekar, 2024). But even with this advantage, tremendous gaps in research remain in AI-driven performance reviews. Specifically, AI models struggle to measure non-numeric traits such as teamwork, leadership, and flexibility (Kellogg et al., 2020). They also struggle with perceiving the interpersonal dynamics at work and the ability to replicate human judgments' empathy and emotional competence (Mehrabi et al., 2021).

Additionally, the questions of equity, ethics, and employee faith in AI-generated decisions need more research into productive applications of AI in performance appraisal systems without prejudicing justice and transparency.

## 2. Literature Review

We have witnessed the extensive integration of artificial intelligence technologies into Human Resource operations, ranging from recruitment and employee engagement to performance management. AI has the potential to drastically improve the efficiency of such processes by automating standard procedures, recognizing patterns in the performance of employees, as well as forecasting future results. Most AI systems, however, still face challenges in evaluating the comprehensive set of characteristics influencing the performance of employees.

A key limitation in AI-based performance evaluation lies in evaluating immeasurable skills like interpersonal communication, conflict management, and teamwork. These attributes most often are described as essential to employee effectiveness, and cannot be readily expressed through systematic data or written feedback. However, LLMs are adept at handling large volumes of text-based information, they can't possibly understand the subtle signals needed to measure complex interpersonal relationships, such as how an employee responds to conflict or interacts with difficult coworkers. This disparity requires further research on how AI can be set up to find and assess these soft skills appropriately.

AI models like LLMs lack the emotional intelligence that managers bring to performance ratings. Traditional appraisals typically rely on empathy and knowledge of context, which are essential in ascertaining complex behaviors and attitudes. AI is unable to replicate such human abilities and thus concerns regarding fairness and comprehensiveness in AI-based ratings are raised. Research is needed to analyze how AI systems can incorporate elements such as emotional intelligence and contextual sensitivity to enhance its decision-making.

Lastly, the trust and acceptability of AI-generated performance evaluations remain underexplored. Employees and managers may be skeptical of AI's ability to make fair, accurate assessments, particularly when it comes to evaluating complex human behaviors. Further research is needed to understand how organizations can build trust in AI-driven performance appraisals and integrate these systems into existing HR frameworks without eroding employee morale or organizational culture.

## 3. Methodology

### 3.1. Model Selection

For this study on performance appraisal automation, the data set is fed into GPT 4 (generative pre-trained transformer 4) because of its exceptional capabilities in natural language processing. Its ability to understand both structured and unstructured data makes GPT 4 particularly valuable for automating performance appraisals. GPT-4 can interpret complex text and produce predictive summaries to automate this process.

Most of this analysis is focused on unstructured feedback from employees taken from self-evaluation and performance goals associated with the employee. The structured data comes from the goals set as a team for that employee

### 3.2. Data Collection

The dataset for this study comprises two primary data sources: unstructured feedback from employee self-evaluations and structured data from mid-year performance reviews and check-ins.

- **Unstructured Feedback:** This dataset is a text file containing employee personal insights gathered from their self-evaluation, which also includes achievements, challenges, and stretch goals undertaken during the performance year. This dataset also includes the personal goals set by the employee at the beginning of the year. This dataset results in a substantial volume of unstructured data due to each record being manually input

and is prone to a lot of errors, including but not limited to missing values, spelling errors, omitting specific evaluation points used as a key performance indicator (KPI).

- **Structured Data:** This dataset is a metric used to set team-level goals that are largely uniform across a set of team members and metrics that assess individual growth, predictive scores for a raise, and a probability of promotion. This dataset is mainly standardized, and hence the structured data is populated on each record, omitting null values and ensuring completeness and comparability across the dataset. This data set can also include attendance records and productivity metrics to help with performance appraisal.

This combination of structured and unstructured data helps robust performance evaluations to mitigate information loss in qualitative assessments (Doshi-Velez & Kim, 2017).

### 3.3. Preprocessing and Data Cleansing

As part of data cleansing and preprocessing, the datasets undergo a rigorous cleanup to ensure compliance with privacy standards by removing confidential information. Unstructured text values and missing elements are eliminated to enhance data quality. The data is tokenized for fields containing numerical metrics, to prepare for input to GPT-4.

The redundant information in the feedback is removed as part of text normalization, text normalization helps remove irrelevant information to preserve context and limit evaluation. The personal identifiers are eliminated to reduce bias and maintain privacy for fair assessments (Mehrabi et al., 2021). Missing values were then handled by applying interpolation techniques to leverage contextual natural language processing for missing qualitative feedback (Nguyen et al., 2023).

After preprocessing, the data was fed into GPT-4, which analyzed both the structured and unstructured data and generated performance appraisals. These appraisals were based on employee goals, key achievements, and feedback from self-evaluations and mid-year reviews.

### 3.4. Model Selection and Data Fine-Tuning

To generate the model, GPT-4 was selected due to its strong natural language processing capabilities, along with its ability to read text and create humanized assessments (Brown et al., 2020). When compared with Llama 2 and Claude, GPT-4 leads due to its optimization capabilities for enterprise applications, which include summarization and structured text generation. Llama 2 struggles with long-form coherence and is less reliable for detailed performance reviews, while Claude is strong in conversational AI, but it is not advanced in document processing and structured reasoning.

The model is fine-tuned using historical performance review data taken over the years and human-labeled appraisals to ensure alignment with corporate performance goals. Prompt engineering and fine-tuning the data can improve domain relevance, reduce bias, and help with coherence in appraisals (Raffel et al., 2020).

### 3.5. Performance Appraisal Generation and Validation

The preprocessed data is then fed into GPT-4 to identify key performance trends and generate a structured summary, including strengths, areas for improvement, and recommendations for professional growth. This is then validated by human evaluators to review AI-generated appraisals and adjust for missing contexts, such as immeasurable skills like interpersonal skills and conflict resolution. The human-in-the-loop process provides feedback to iteratively improve GPT-4 outputs (Lundberg & Lee, 2017). Integration of AI into performance appraisals has the potential to enhance efficiency, reduce bias, and automate the process. However, artificial intelligence algorithms, such as GPT-4, tend to struggle in evaluating live workplace behaviors, career development patterns, and issues encountered by employees. Human-in-the-loop (HITL) workflows ensure that human reviewers read AI-generated reviews, offering perceptions that may elude artificial intelligence. An employee's underperformance, for example, may be caused by extraneous circumstances such as personal issues or company restructuring, which cannot be fully understood by artificial intelligence (Wu et al., 2021).

It is essential to consider the human-in-the-loop (HITL) approach because it involves dynamic collaboration where humans and algorithms iteratively influence each other, leading to continuous model improvement (Wu et al., 2021). By TL into performance appraisal systems can mitigate biases in automated evaluations and ensure individual nuances and contextual factors are considered this hybrid approach helps computational efficiency while maintaining the depth of human judgment (Wu et al., 2021). As AI models can amplify biases present in training data, the human-in-the-loop approach allows HR specialists to identify and correct potential biases, ensuring fairness.

Although the human-in-the-loop approach significantly enhances the robustness of AI-driven automated performance appraisals, several challenges continue to remain. It is important to develop efficient workflows that streamline and limit human involvement (Andersen & Maalej, 2023). Large organizations may struggle to allocate sufficient human reviewers to validate AI-generated reviews, although the cost of hiring HR professionals to provide ongoing feedback to AI systems can reduce the economic benefits of automation (Wu et al., 2021). Human intervention in data annotation, validation, and bias correction as the volume of data increases, and the number of human touch points needed for oversight expands, creating bottlenecks that slow down the evaluation process and limit the efficiency that AI is meant to offer (Wang et al., 2022).

To address these scalability issues, organizations can implement strategies such as selective human oversight or conduct semi-automated decision support along with active learning techniques. Instead of reviewing every AI-generated appraisal, human evaluators can focus on cases with lower AI confidence levels. These reviews can be prefiltered using rule-based systems to ensure minimal human intervention (Andersen & Maalej, 2023).

### 3.6. Deployment and Continuous Improvement

GPT-4 can be deployed as an API integrated into HR systems, such as Workday, SAP, or custom-built platforms. Data preprocessing is then done to ensure confidentiality, anonymizing sensitive information before feeding it to GPT-4. AI-generated performance summaries should always be presented as recommendations rather than final decisions and have managers or HR professionals review them to provide feedback or override decisions if necessary. The system should ensure that confidence scores are presented to highlight areas where human review is recommended. Some areas where confidence scores may be low include missed check-ins, null values, and unusual performance trends compared to historic data.

There is always a need for a continuous improvement framework to enable effective ongoing monitoring and feedback-driven refinement. To ensure alignment with organizational goals, regular evaluation of AI-generated appraisals against human-generated assessments is always necessary (Mehrabi et al., 2021). As part of continuous improvement and model optimization, it is imperative to ensure a structured review process where inaccuracies and missing contextual nuances can be flagged. Feedback loops should allow the model to learn from expert judgments while maintaining compliance with ethical AI guidelines (Raji et al., 2020).

## 4. Results

The use of GPT-4 for automating performance evaluations has demonstrated promising outcomes, including strengths and limitations that can be improved to enhance its effectiveness in organizational settings.

GPT-4's ability to generate accurate and consistent performance evaluations has been proven to be one of the model's strengths. This model has demonstrated proficiency in processing the combination of structured and unstructured feedback to ensure consistency reflected in the appraisals. Employees with similar performance characteristics received comparable assessments, which contrasts with traditional human evaluations and can vary between different reviewers, even for similar performance levels. By leveraging this consistency and applying predefined performance criteria, GPT-4 provides a standardized assessment that reduces appraiser variability (Liu et al., 2020).

A huge amount of time was saved by automating the performance appraisal process. GPT-4's ability to quickly process large datasets and generate reports has enabled HR professionals to spend more time on making strategic decisions and helping employee development rather than spending time on individual appraisals. This model efficiently analyzed large amounts of data to produce appraisal results in a fraction of the time that would have been spent on manual processing (Binns et al., 2018).

Although the results generated by GPT-4 were largely accurate and aligned well with expected outcomes, a few appraisals lacked the depth of empathy present in human-generated evaluations. Although these are qualitative, but not quantitative skills, it is essential to consider these factors since they play a major role in obtaining a raise or promotion for the employee. These nuanced aspects are always difficult for artificial intelligence models to fully capture because of the lack of inherent human empathy for subjective assessment.

These limitations underscore the importance of the human-in-the-loop approach, where human evaluators define benchmarks, override the recommendation given by the AI-based approach, and ensure that these appraisals are always aligned with the dynamic nature of employee performance and incorporate organizational standards to ensure fair and free of unintended discrimination (Gupta & Tembhurnekar, 2024).

The ability to process and extract meaningful insights from unstructured feedback was not highly optimistic with GPT-4, particularly with qualitative comments within the data. This limitation indicates a need to incorporate sentiment analysis and better handle unstructured data to improve model assessment for non-technical attributes (Li & Li, 2021).

## 5. Conclusion

Large language models (LLMs) such as GPT-4 are known to significantly enhance the performance appraisal process by minimizing the need for manual reviews, which in turn saves time. This automation technique for performance evaluations using large language models, particularly GPT-4, helps increase efficiency by enhancing pattern recognition using a data-driven approach to assess employee performance. GPT-4 demonstrates significant strengths in its ability to analyze both structured and unstructured feedback to provide consistent evaluations and automate the process efficiently. Organizations can therefore leverage AI techniques to streamline performance reviews and reduce subjective biases that occur in traditional evaluations.

Despite the strong benefits this research highlights several limitations of LLM-based appraisals, such as the lack of human empathy and the inability to handle missing data, which remain critical concerns. GPT-4 struggles with assessing interpersonal skills, such as teamwork, conflict resolution, and emotional intelligence, which are some of the crucial components of employee performance that are difficult to quantify through AI models alone. These limitations highlight the importance of human-in-the-loop approaches, where human oversight is often integrated with AI-generated appraisals, which adjust for bias to ensure fairness.

There is also a need for industry-specific algorithms tailored for performance evaluations that integrate sentiment analysis and effective computing to better interpret feedback and improve techniques for handling missing data. However, even with all these integrations, there is a need for continuous monitoring and iterative refinement of performance appraisal systems.

## References

[1] DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? Journal of Applied Psychology, 102(3), 421-433.

[2] Aguinis, H., Joo, H., & Gottfredson, R. K. (2019). Performance management universals: Thinking globally and acting locally. *Annual Review of Organizational Psychology and Organizational Behavior, 6*(1), 305-331.

[3] Giles, W., Kianto, A., & Van Zyl, L. (2021). Performance appraisals in modern organizations: A review and future research agenda. Human Resource Management Journal, 31(4), 678-698.

[4] Gupta, R. K., & Tembhurnekar, C. M. (2024). AI-driven performance appraisal systems: A critical literature review of emerging issues and challenges. ShodhKosh: Journal of Visual and Performing Arts, 5(7), 492-496.

[5] Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366-410.

[6] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

[7] Nyathani, R. (2023). AI in performance management: Redefining performance appraisals in the digital age. Journal of Artificial Intelligence & Cloud Computing, 1(2), 134-145.

[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS).

[9] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[10] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpretable machine learning predictions. Advances in Neural Information Processing Systems.

[11] Nguyen, T., Kremer, S., & Shawe-Taylor, J. (2023). Handling missing data in AI applications: A review of current techniques. Journal of Machine Learning Research.

[12] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research.

[13] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2021). A survey of human-in-the-loop for machine learning. arXiv preprint arXiv:2108.00941.

[14] Andersen, J. S., & Maalej, W. (2023). Design patterns for machine learning-based systems with human-in-the-loop. arXiv preprint arXiv:2312.00582.

[15] Wang, J., Guo, B., & Chen, L. (2022). Human-in-the-loop machine learning: A macro-micro perspective. arXiv preprint arXiv:2202.10564.

[16] Raji, I. D., Mitchell, M., Smith, J., & Kumar, I. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. ACM FAT Conference. https://arxiv.org/abs/2001.00973

[17] Binns, R., & Creese, S. (2018). Reducing algorithmic bias in recruitment: A case study. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 415-426.

[18] Liu, Y., Zheng, X., & Tang, L. (2020). An AI approach to employee performance evaluation. Journal of Business Research, 112, 103-115.