Rilke, Rainer Michael; Sliwka, Dirk

**Working Paper**

# When algorithms rate performance: Do large language models replicate human evaluation biases?

ECONtribute Discussion Paper, No. 384

**Provided in Cooperation with:**
Reinhard Selten Institute (RSI), University of Bonn and University of Cologne

*Suggested Citation:* Rilke, Rainer Michael; Sliwka, Dirk (2026) : When algorithms rate performance: Do large language models replicate human evaluation biases?, ECONtribute Discussion Paper, No. 384, University of Bonn and University of Cologne, Reinhard Selten Institute (RSI), Bonn and Cologne

This Version is available at:
https://hdl.handle.net/10419/339347

ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 384

# When Algorithms Rate Performance: Do Large Language Models Replicate Human Evaluation Biases?

Rainer Michael Rilke Dirk Sliwka

January 2026 (updated version)

www.econtribute.de

UNIVERSITÄT BONN

UNIVERSITÄT ZU KÖLN

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 – 390838866 – is gratefully acknowledged.

# When Algorithms Rate Performance: Do Large Language Models Replicate Human Evaluation Biases?*

Rainer Michael Rilke† Dirk Sliwka‡

January 12, 2026

A large body of research across management, psychology, accounting, and economics shows that subjective performance evaluations are systematically biased: ratings cluster near the midpoint of scales and are often excessively lenient. As organizations increasingly adopt large language models (LLMs) for evaluative tasks, little is known about how these systems perform when assessing human performance. We document that, in the absence of clear objective standards and when individuals are rated independently, LLMs reproduce the familiar patterns of human raters. However, LLMs generate greater dispersion and accuracy when evaluating multiple individuals simultaneously. With noisy but objective performance signals, LLMs provide substantially more accurate evaluations than human raters, as they (i) are less subject to biases arising from concern for the evaluated employee and (ii) make fewer mistakes in information processing closely approximating rational Bayesian benchmarks.

**Keywords**: Performance Evaluation, Large Language Models, Signal Objectivity, Algorithmic Judgment, Gen-AI

**JEL** Codes: J24, J28, M12, M53

*Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 – 390838866.

†WHU - Otto Beisheim School of Management, Economics Group, Burgplatz 2, 56179 Vallendar, Germany. rainer.rilke@whu.edu

‡University of Cologne, MPI for Behavioral Economics, CESifo, IZA, Faculty of Management, Economics, and Social Sciences, Albertus-Magnus-Platz, D-50923 Köln, Germany. dirk.sliwka@uni-koeln.de

## Introduction

In most organizations the performance of employees is assessed subjectively by supervisors. A large literature in management, economics, accounting, and psychology has consistently documented systematic patterns in these performance evaluations: the lowest evaluation categories are rarely used even when available; ratings cluster heavily around the midpoint of evaluation scales, a phenomenon known as *centrality bias*; and substantially more employees receive ratings in the upper range than the lower range (often called *leniency bias*) (see e.g. Landy and Farr 1980; Bretz Jr, Milkovich, and Read 1992; Murphy and Cleveland 1995; Jawahar and Williams 1997; Prendergast and Topel 1996; Prendergast 1999; Moers 2005; Bol 2008; Golman and Bhatia 2012). Such biases may impose significant costs on organizations by distorting incentives, misallocating talent, and potentially undermining the credibility of performance management systems.

The adoption of Large Language Models (LLMs) in organizations raises fundamental questions about how these systems affect personnel decsions. On the one hand, organizations are increasingly deploying LLMs for tasks ranging from resume screening to information provision for employee performance assessment, sometimes with the implicit assumption that algorithmic evaluation might reduce human biases. On the other hand, the use of AI in general to assess the performance of humans in organizations is also often controversially discussed, causing fear of "big brother"-type of supervision or even of creating new biases and lack of procedural fairness or transparency in the rating process (Tambe, Cappelli, and Yakubovich 2019; Kantor and Sundaram 2022; Bol, Brown, and LaViers 2025). Yet we currently lack systematic evidence on how LLMs actually behave when tasked with evaluating performance and whether and how their evaluation quality depends on the nature of the information being assessed.

This paper documents performance evaluations conducted by LLM across settings that vary in signal objectivity and evaluation format. We conduct three complementary studies progressing from purely subjective judgments (CEO evaluations without objective benchmarks) through semi-objective assessments (job applications with induced quality levels) to noisy but objective performance data (experimental outcomes with clear Bayesian benchmarks).

We first show that without clear objective performance standards, LLMs display the same evaluation patterns commonly observed among human raters. For example, when we prompt an LLM to assess the performance of S&P 500 CEOs using a generic standard five-point scale ("1 = Unsatisfactory" to "5 = Outstanding") commonly used by firms for performance evaluations, the resulting distribution closely mirrors real-world managerial evaluations: the lowest two categories are rarely used, ratings cluster around the midpoint, and there is a pronounced tendency toward leniency, with the top categories assigned far

more frequently than the bottom ones. Strikingly, these patterns persist even when the LLM is explicitly instructed to match a prespecified distribution. When asked to assign the lowest rating to CEOs in the bottom 20% of the performance distribution, the model assigns only about 0.2% of CEOs to this lowest quintile. We then examine whether evaluating multiple CEOs simultaneously mitigates this effect. While joint evaluation increases rating dispersion, the LLM still assigns only 0.8% of CEOs to the bottom quintile when rating groups of five CEOs at once.

We replicate and extend the analysis in a second study where we task the LLM to evaluate job applications for three job entry positions. Here we exogenously induce quality difference by constructing applications of different quality levels. We again find that LLMs exhibit systematic rating leniency when we use the generic rating scale. Greater differentiation is again achieved when multiple applications are evaluated simultaneously and when we prompt the LLM to assign ratings to match a prespecified distribution. Here the LLM is able to achieve this target distribution quite closely when evaluating five applications simultaneously in one prompt.

Finally, we draw on data from a recent experimental study by Kusterer and Sliwka (2024) conducted with crowd workers on Amazon MTurk. In this experiment, subjects completed well-defined tasks and human raters evaluated them using noisy but objective performance signals. As Kusterer and Sliwka (2024) show, human raters display substantial leniency when they know their evaluations determine a bonus for the worker. We re-evaluate each worker from this experiment using an LLM, providing the model with exactly the same information available to the human raters. The LLM's evaluations substantially outperform those of human raters: they are far more accurate and exhibit no leniency. Indeed, the LLM's ratings come remarkably close to the Bayesian-optimal benchmark that can be achieved by combining all available information, and they do so regardless of whether the model is informed that its rating affects the worker's bonus.

Our findings make several contributions to the literature. First, we show that, when there are no objective performance standards and the LLM is prompted to evaluate performance of individuals separately, LLM ratings exhibit the classical patterns of leniency and centrality found by numerous studies on human raters. Second, we provide evidence on LLM evaluation behavior across different rating formats and levels of information objectivity. Our finding that LLM ratings acheive higher dispersion and accuracy when the LLM evaluated more than one individual in one prompt mirrors findings from psychology showing that comparative performance evaluations by human raters lead to more dispersed and more accurate ratings (Heneman 1986; Wagner and Goffin 1997; Becker and Miller 2002; Goffin and Olson 2011). Our paper also contributes to a literature in economics and accounting on the informativeness of subjective performance evaluations (Prendergast and Topel (1996), Bol (2008), Moers (2005), Manthei and Sliwka (2019)). Prendergast and Topel (1996), for

instance, develop a formal model of performance evaluations where supervisors receive noisy signals about employees' performance and trade off preferences for rating accuracy (which induces the aim to use the available signals in a rational Bayesian way) with social concerns and favoritism towards the rated employee, leading to biased evaluations. Indeed, recent experimental work (Kusterer and Sliwka 2024; Ockenfels, Sliwka, and Werner 2025) shows that *on average* across a population of human raters such a Bayesian framework can organize observed human evaluation behavior quite well.[1] But the experiments also reveal that there is substantial noise such that ratings performed by individiual raters differ substantially from optimal Bayesian information processing benchmarks. In other words, human rating errors are much larger than rational information processing would predict. Our results show that when noisy but objective performance information is available, an LLM can produce much more accurate ratings than human evaluators and do so for two reasons: It is (i) less prone to favoritism or social concerns for the ratee and (ii) more able to rationally use the available information to perform Bayesian updating reducing cognitive limitations.

Our study also connects to the growing literature on the use of AI and algorithms in labor markets in general. Much of this work has examined how algorithmic tools affect hiring outcomes by assisting recruiters in screening applicant information prior to interviews or in evaluating candidates afterward (Horton 2017; Avery, Leibbrandt, and Vecci 2024; Li, Raymond, and Bergman 2025; Dargnies, Hakimov, and Kübler 2024; Jabarian and Henkel 2025). As pointed out in this literature, AI can be useful by either replicating human evaluations at lower costs or by also improving rating accuracy. Our results provide evidence in line with the view that LLM may not only reduce costs, but can provide more accurate evaluations when noisy but objective performance information is available. Finally, our findings relate to the broader literature on how machine learning and LLMs can improve human decision making (Kleinberg et al. 2018; Dietvorst, Simmons, and Massey 2018; Ludwig and Mullainathan 2024; Mullainathan 2025). While this work typically studies algorithms as decision aids that reduce noise or bias in human judgments, we examine a setting in which the LLM itself acts as the evaluator. By showing that LLMs replicate human rating biases when objective standards are absent, yet approach Bayesian-optimal accuracy when objective but noisy signals are available, we identify the conditions under which LLMs can meaningfully outperform human raters.

[1] For instance, when regressing evaluations on observed signals the estimated function is remarkably close to evaluation behavior a fully rational decision maker (who also cares for the payoff of the evaluated worker) would show.

## Common Performance Evaluation Patterns

Before reporting our results, we briefly review common patterns in human performance evaluations documented in the prior literature as a benchmark for interpreting LLM behavior in our studies. In Figure 1 we plot the distribution of performance ratings within firms using 5-point evaluation scales reported in published studies.[2] The figure illustrates several well-documented regularities: (i) even when rating systems provide five levels, the lowest two levels are very rarely used; (ii) the largest proportion of employees is assigned to the middle category or one level above it, a pattern often referred to as *centrality bias* or *central tendency*; and (iii) substantially more employees are evaluated at the upper levels than at the lower ones, a phenomenon commonly termed *leniency bias*.

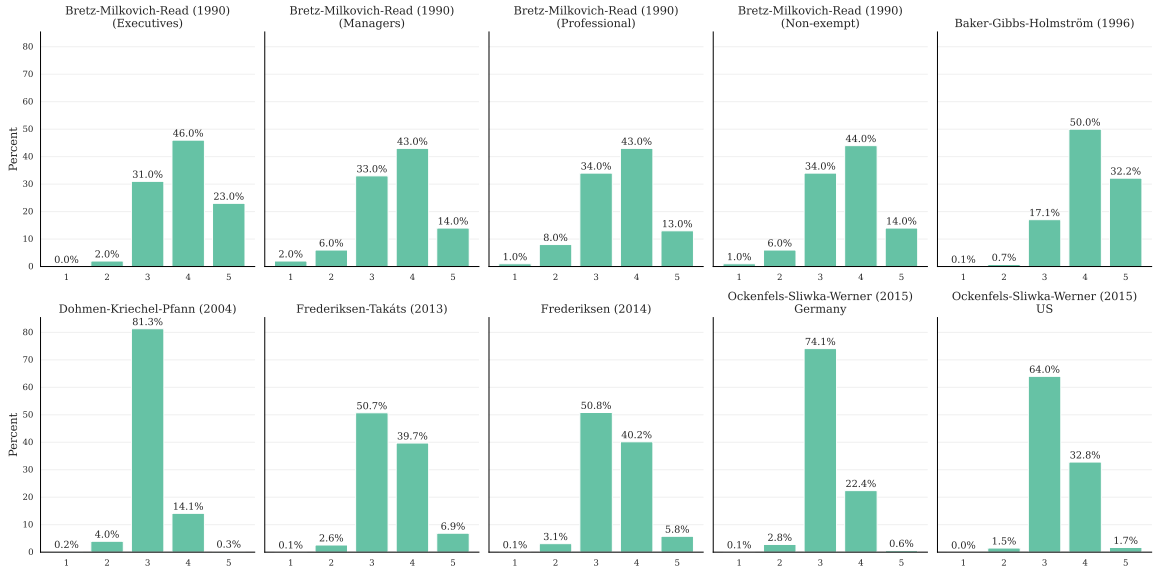



Figure 1: Distribution of Performance Ratings across Various Studies

## Method

Our approach systematically documents LLM evaluation behavior by eliciting performance ratings through structured prompts. Across all three studies, we maintain a consistent technical protocol while systematically varying the format in which evaluations are elicited.

In each study, we construct evaluation prompts that specify (i) the subject to be evaluated (CEO, job application, or worker performance), and (ii) the performance information avail-

[2] Bretz Jr, Milkovich, and Read (1989) report rating distributions collected from a survey among 63 of the Fortune 100 companies at the time. The distributions from Baker, Gibbs, and Holmstrom (1994), Dohmen, Kriechel, and Pfann (2004), Flabbi and Ichino (2001), Frederiksen (2013), Frederiksen and Takáts (2011) have been extracted from Frederiksen, Lange, and Kriechel (2017). Ockenfels, Sliwka, and Werner (2015) report performance ratings in a multinational firm with 3,822 managers in Germany and 2,540 in the US.

able to the evaluator. All prompts are submitted to GPT-5-mini via OpenAI's API. The model is instructed to return only a numerical rating, and we employ robust parsing procedures. This approach ensures comparability across studies while allowing us to systematically manipulate key features of the evaluation task.

A central design feature across all studies is the systematic variation between individual and comparative evaluation formats. In individual evaluation, the LLM receives information on a single subject per prompt and provides one rating. In comparative evaluation, the LLM receives information on multiple subjects simultaneously (either 3 or 5) and provides a rating for each. All subjects are evaluated under multiple formats, enabling within-subject comparisons of how evaluation context shapes judgments.

The three studies are ordered by the degree to which objective performance benchmarks exist. Study 1 (CEO evaluations) represents subjective judgment with no fully objective ground truth. In Study 2 (job applications) we actively manipulate quality: we use GPT-5-mini to generate application texts with pre-assigned quality levels, then have a separate LLM instance evaluate them without knowledge of induced quality difference. Study 3 (objective performance) employs experimental data from a recent study by Kusterer and Sliwka (2024) with known performance distributions and observable performance signals, providing a clear Bayesian benchmark for rational evaluation. Moreover, here we can compare LLM evaluations to those made by human raters who had evaluated performance in the original experiment. This progression allows us to examine whether—and how—evaluation quality varies with the objectivity of available information. We employ two types of rating scales across studies. The *generic scale* use qualitative labels (e.g., "Outstanding," "Meets Expectations") that provide no explicit guidance about expected rating distributions. The *distribution scale* explicitly anchor ratings to percentile ranges (e.g., "5 = top 20%"), providing clear guidance about how ratings should be distributed akin to Forced/or Recommended Distributions in firms (see e.g. Berger, Harbring, and Sliwka 2013; Cardinaels and Feichter 2021; Bond 2025). This variation allows us to examine whether explicit distributional guidance affects LLM rating behavior, paralleling common practices in organizational performance management. Finally, in study 3 we also prompt the LLM to rate an objective performance outcome on a continuous scale to directly compare evaluations to a Bayesian standard.

## Study 1: Evaluating CEOs

### Independent evaluations

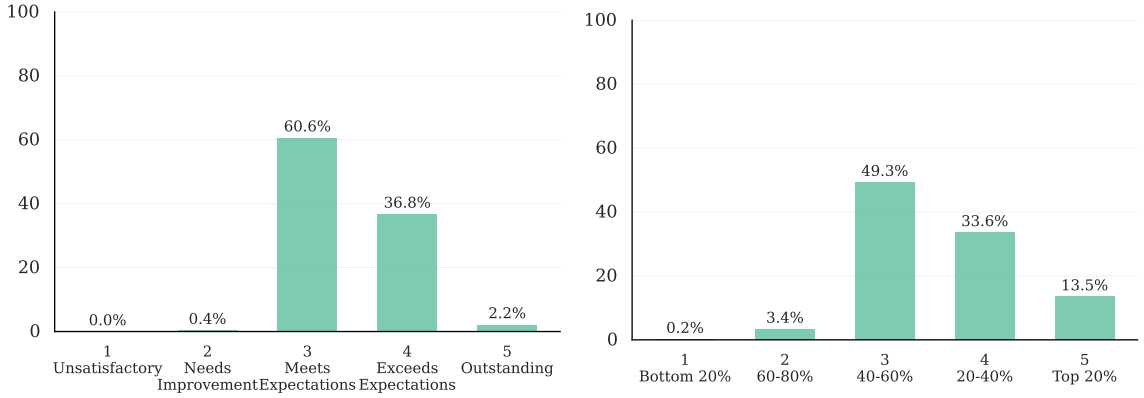
In a first study we tasked an LLM to evaluate the performance in 2024 of all CEOs of the S&P 500 firms following on a standard 5 point evalation scale. We used two different prompts and respective rating scales reflecting typical appraisal practices within firms: We start with

an often used rather generic rating scale promptig the LLM as follows:

> *Rate the performance of the CEO {ceo} of the company {company} in 2024 on a scale of 5 to 1:*
>
> *5 = Outstanding / Exceptional*
>
> *4 = Exceeds Expectations*
>
> *3 = Meets Expectations*
>
> *2 = Needs Improvement*
>
> *1 = Unsatisfactory*

Of course, with such a scale it is impossible to clearly identify biases as there is no entirely objective definition of whether a manager's performance was "exceptional" or "needs improvement". To give more guidance some firms provide recommended distributions by specifying what share of managers should receive which ranking. We thus also collect evaluations providing such distributional guidance. Specifically we now use the following prompt:

> *Rate the performance of the CEO {ceo} of the company {company} in 2024 on a scale of 5 to 1. Award a 5 if you think {ceo} is among the best 20% CEOs in the S&P 500 companies. Award a 4 accordingly if {ceo} is among the best 40% but not the best 20% (2nd quintile), and continue analogously according to the quintiles so that you award a 1 if {ceo} is among the worst 20%:*
>
> *5 = among the best 20% of S&P 500 CEOs*
>
> *4 = best 40% but not best 20% (2nd quintile)*
>
> *3 = middle 20% (3rd quintile)*
>
> *2 = bottom 40% but not bottom 20% (4th quintile)*
>
> *1 = among the worst 20%*




(a) Generic Scale

(b) Distribution Scale

Figure 2: Distribution of CEO Ratings by Scale Type (Precentage of CEOs)

The key results are displayed in Figure 2 which shows the distribution of the assigned ratings. Our first key observation is that we observe evaluation patterns that are very similar to those typically reported in studies of subjective performance evaluations by human raters within firms as reported in section Section : For one, there is subsantial "leniency" as the lowest possible rating is virtually never assigned, and the second lowest rating is also extremely rare. Moreover, there is a tendency for "centrality" as for instance in the generic scale (left panel of Figure 2) about 60% of all managers are evaluated with the "meets expectations" midpoint. When using the distribution scale the variance of ratings slightly increases but only for above average ratings. The LLM here received the clear instruction to identify to which performance quintile a person belongs and it clearly fails in this task here – just as human evaluators commonly do when faced with the same evaluation task. Only 0.2% of all CEOs are rated into the bottom quintile (rating = 1) and only 3.4% into the second quintile despite the fact that by construction 20% of all CEOs should be assigned to each of these levels.

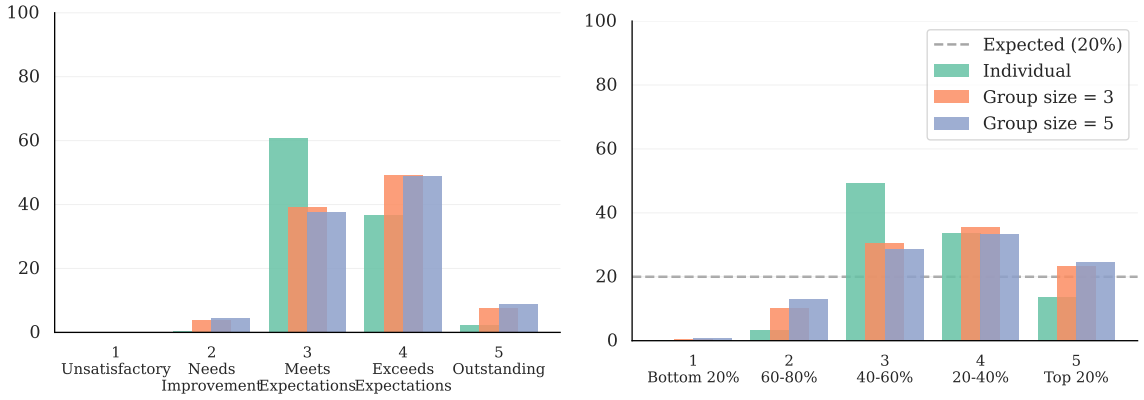
Finally, the use of the distribution scale somewhat mitigates the centrality observed in the generic scale as the share of CEOs evaluated at the modpoint drops and the variance of ratings increases (Levene's test, $p<0.001$). But rating dispersion is still rather limited. Moreover, rather than reducing leniency, the use of the distribution scale increases average ratings (t-test, $p<0.001$) as it increases the likelihood of assigning the two top ratings more than that of assigning the two lowest rating categories.

### Joint evaluations

So far the LLM had been tasked to evaluate each CEO independently in a separate prompt, i.e. mimicking a situation where a supervisor has to evaluate a single employee. We next

examine whether the task becomes easier for the LLM when it evaluates multiple managers simultaneously rather than one at a time. Prior work in psychology and management shows that comparative performance evaluations tend to yield more dispersed and more accurate ratings (Heneman 1986; Wagner and Goffin 1997; Becker and Miller 2002; Goffin and Olson 2011). To assess whether this holds for LLMs, we randomly assign managers to groups of three or five and prompt the model to rate the performance of each CEO within these groups. Instead of evaluating a single CEO per prompt, the LLM now receives a list of three or five CEOs and is asked to rate them jointly:

> *Rate the performance of the following CEOs in 2024 on a scale of 1 to 5. Award a 5 if you think the CEO is among the best 20% CEOs in the S&P 500 companies. Award a 4 accordingly if the CEO is among the best 40% but not the best 20% (2nd quintile), and continue analogously according to the quintiles so that you award a 1 if the CEO is among the worst 20%. Answer only with the number 1, 2, 3, 4 or 5 for each CEO:*




(a) Generic Scale

(b) Distribution Scale

Figure 3: Distribution of CEO Ratings by Scale Type Across Group Sizes (Precentage of CEOs)

The results are shown in Figure 3. Indeed, when the LLM evaluates managers in groups, ratings are more differentiated for both scale types. Figure 4 shows the standard deviation of ratings by group size and scale type. Levene's test rejects the null hypothesis of equal variances when moving from individual to group evaluations both for the generic and the distribution scale (see bar spanners in Figure 4 for respective p-values). Rating dispersion increases slightly further when moving from groups of three to groups of five, but here we cannot reject the Null of equal variances. Also note that the distribution scale leads to significantly more differentiation than the generic scale for each group size.

Interestingly, we still observe a strong "reluctance" of the LLM to assign the lowest rating of "1" in all 6 different settings. Even when evaluating in groups of five under the distribution

scale, only 0.80%. of all CEOs are with "1" whereas about 20% should be "1" if the LLM would perfectly follow the prompt instructions.

A natural explanation for the persistent leniency observed in our CEO evaluations is that LLMs inherit systematic patterns present in the human-generated texts on which they are trained. Because performance evaluations, managerial communication, and public discourse about leaders tend to be positively skewed and often avoid extreme negative judgments, the model has learned linguistic priors that favor moderate or favorable assessments. When objective standards are absent and the model must rely on these learned priors, its ratings mirror the leniency and centrality biases documented among human evaluators. Evaluating multiple individuals simultaneously mitigates this tendency somewhat, because the model can anchor its judgments on relative differences within the group rather than relying solely on its learned prior distribution. Comparative evaluation thus provide a counterweight to the model's inherited positivity bias, enabling more differentiation even though the underlying reluctance to assign very low ratings persists.

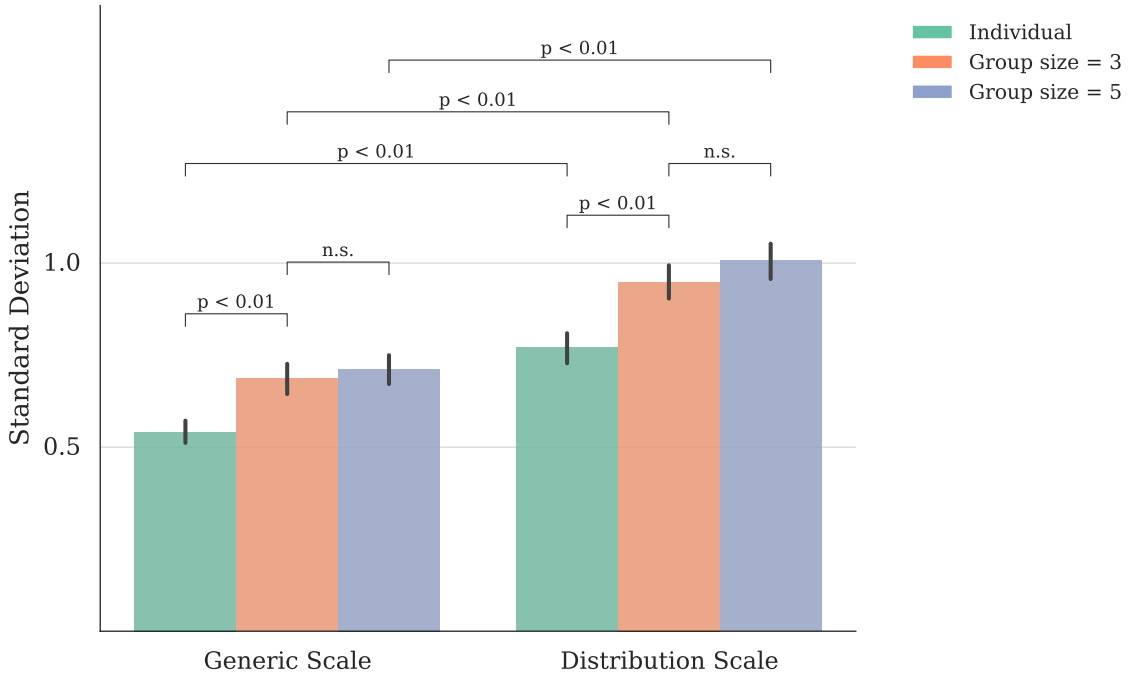



Figure 4: Standard Deviation of CEO Ratings by Scale Type and Group Size

Table 1: The Effect of Scale Type and Group Size on CEO Ratings

| | Rating | | |
|---|---|---|---|
| | Generic Scale | Distribution Scale | All |
| | (1) | (2) | (3) |
| Group size=3 | 0.198*** | 0.144*** | 0.171*** |
| | (0.027) | (0.033) | (0.023) |
| Group size=5 | 0.214*** | 0.108*** | 0.161*** |
| | (0.028) | (0.035) | (0.024) |
| Distribution Scale | | | 0.106*** |
| | | | (0.021) |
| Intercept | 3.407*** | 3.567*** | 3.434*** |
| | (0.024) | (0.035) | (0.023) |
| Observations | 1,503 | 1,503 | 3,006 |
| $R^2$ | 0.022 | 0.004 | 0.014 |

Robust standard errors clustered on the CEO in parenthes. * p < 0.1, ** p < 0.05, *** p < 0.01

Table 1 shows a regression of the rating on the group size and the scale type to analyze how group evaluations shift rating leniency. While moving to evaluation in groups increases rating dispersion it does not reduce rating leniency as the table shows. On the contrary, average ratings increase when moving from individual to group evaluations. Similar to the effect of using the distribution scale, the increase in disperson thus does not come long with lower leniency. On the contrary, ratings become even more positive on average as the frequency of the two top ratings increases more than that of the two lowest rating categories.

## Study 2: Evaluating Job Applications

In the preceding section, we have shown that LLM ratings of CEO performance exhibit very similar rating patterns as human raters within organizations. However, a key limitation of this setting is the absence of a strict objective benchmark for performance.[3] As a result, it is impossible to determine whether observed biases reflect failures of the LLM to accurately assess performance or simply inherent challenges in evaluating CEO performance subjectively. To move closer to having a clearer benchmark, we conducted a second study in which we exogenously induce quality differences. To do this, we first use an LLM to

[3]For instance, there is no clear objective standard for key trade-offs such as favoring current profitability versus future value generation, or financial performance versus societal impact etc.

generate job application texts with pre-assigned quality levels, then have a separate LLM instance evaluate them without knowledge of induced quality.

For three job titles (Junior Business Analyst, Graduate Engineer Trainee, and Junior Software Developer), we prompt the LLM to write 200 words job application texts of varying underlying quality. The quality level for each text is determined by a random draw from a uniform distribution with five quality levels instructing the LLM for instance to generate an application belonging to the bottom 20% of applications when the drawn quality level is equal to 1.[4]

In a separate stage, another LLM instance (acting now as evaluator) rates each application's quality. The model receives the job title and application text and is instructed to rate the application. As in the CEO study, we employ both rating scales: the Generic Scale with qualitative labels ("Outstanding," "Exceeds Expectations," "Meets Expectations," "Needs Improvement," "Unsatisfactory") and the Distribution Scale with explicit percentile anchoring. For the Distribution Scale, the evaluation prompt reads:

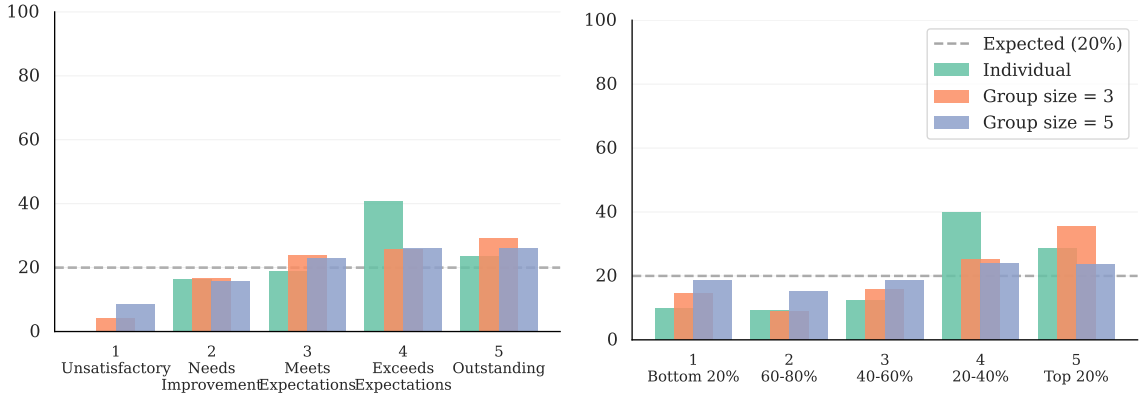
> *Rate the quality of the following application(s) on a scale of 5 to 1. Award a 5 if you think the application is among the top 20% of applications. Award a 4 accordingly if the application is among the top 40% but not the top 20% (2nd quintile), and continue analogously according to the quintiles so that you award a 1 if the application is among the bottom 20%.*
>
> *5 = among the top 20% of applications*
>
> *4 = top 40% but not top 20% (2nd quintile) of applications*
>
> *3 = middle 20% (3rd quintile) of applications*
>
> *2 = bottom 40% but not bottom 20% (4th quintile) of applications*
>
> *1 = among the bottom 20% of applications"*
>
> *Answer only with the number 5, 4, 3, 2, or 1.*

As in the above, we elicit ratings in different group sizes ranging from one application at a time to groups of three and five applications randomly drawn from all generated applications for the same job. Hence, each application is evaluated multiple times under different evaluation formats (individual vs. groups of 3 vs. groups of 5) and both rating scales (Generic vs. Distribution).

The key results are displayed in Figure 5. Again, as shown in panel (a), under the generic evaluation scale we observe persistent leniency effects: the LLM underutilizes the lowest categories and concentrates ratings in the upper range producing distributions right-skewed relative to the expected uniform benchmark.[5] The dominant pattern here—stronger leniency

[4]The template of the application generation prompt and further information can be found in the Appendix.

[5]Figure A1 reported in the Appendix shows confusion matrix plots to provide more details on the specific drivers of this leniency effects showing the distribution of ratings for applications of each quality level separately.




(a) Generic Scale

(b) Distribution Scale

Figure 5: Distribution of Job Application Ratings by Scale Type Across Group Sizes (Percentage of Applications)

with less centrality—differs from Study 1 (CEO evaluations), where ratings exhibited stronger centrality alongside more moderate leniency.[6]

When using the Distribution Scale, the LLM produces a rating distribution that is much closer to the expected uniform benchmark. Notably, when evaluating applications in groups of five, the LLM closely approximates the target distribution, assigning approximately 20% of applications to each rating category. This finding contrasts with Study 1 (CEO evaluations), where even group evaluations under the Distribution Scale failed to achieve the target distribution. The ability of the LLM to align ratings with specified distributions in this semi-objective setting suggests that clearer performance benchmarks facilitate more differentiated evaluations.

Figure 6 shows the standard deviation of ratings for the different conditions reports p-values of paired Wilcoxon signed-rank tests that compare the absolute deviations from treatment-specific means across evaluation formats for the same applications. This paired design accounts for application-specific heterogeneity by testing whether variance differs within applications across formats. The tests reject the null hypothesis of equal variances when moving from individual to group evaluations for both scales ($p<0.01$), showing again that comparative evaluation significantly increases rating differentiation. The Distribution

Under individual evaluation with the Generic Scale, the LLM rarely assigns low-quality applications (quality 1–2) to applications from the respective induced quality categories. Instead, low-quality applications are systematically shifted upward: quality-1 applications are predominantly rated 2 or 3 rather than 1, while quality-2 applications cluster around rating 3 and 4. High-quality applications (quality 4–5) are more accurately identified, but the overall effect is substantial rating inflation across the quality distribution.

[6]Note that context-specific variation in the dominance of either leniency or centrality is also observed in firm level settings where the distribution reported in some firms show predominantly central tendency (e.g., Dohmen, Kriechel, and Pfann (2004); Ockenfels, Sliwka, and Werner (2015)), while others exhibit more pronounced leniency (e.g., Baker, Gibbs, and Holmstrom (1994)) (see Figure 1).

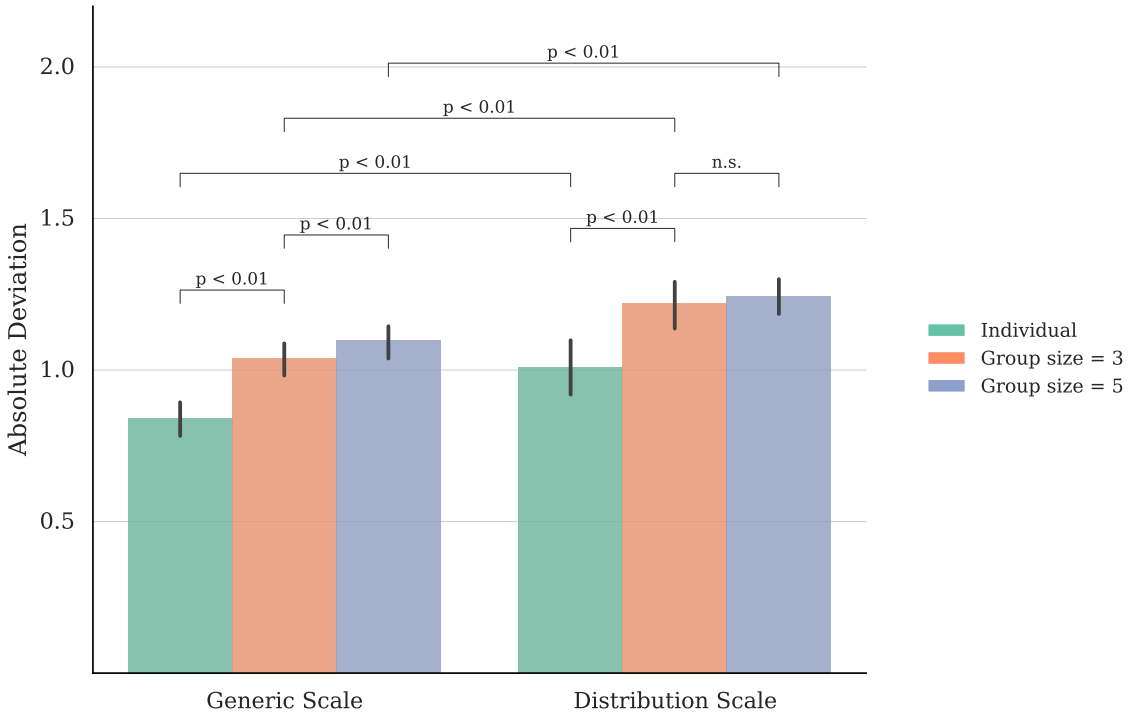



Figure 6: Deviation between Assessed and Induced Quality of Job Applications by Scale Type and Group Size

Scale also produces greater dispersion than the Generic Scale across all group sizes (all comparisons p<0.01), reflecting its explicit percentile guidance.

As Table 2 shows–which reports regressions of assigned ratings group size, and scale type controlling for application fixed effects–not only dispersion is increased here but at the same time leniency is reduced when the LLM evaluates applications in groups rather than individually. Note that this contrasts with the results obtained in the CEO study above, where the increase in dispersion came along with even more leniency. This suggests that group evaluation effects depend on the initial bias structure: when individual evaluations exhibit strong centrality (CEO study), comparative judgment increases differentiation by permitting more extreme ratings in both directions, with leniency dominating. When individual evaluations exhibit stronger leniency at the outset (job applications), comparative judgment enables downward correction as the LLM more easily recognizes inflated assessments relative to peers.

Finally, we can now make use of the fact that we induced different quality levels to study the effect of the conditions on the difference between assessed and induced quality as an inverse measure of rating accuracy. Table 3 reports regressions on the absolute difference between the rating and application quality on group size and scale type controlling again

Table 2: The Effect of Scale Type and Group Size on Job Application Ratings

| | LLM Rating | | |
|---|---|---|---|
| | Generic | Distribution | Pooled |
| | (1) | (2) | (3) |
| Group size = 3 | -0.126*** | -0.098*** | -0.112*** |
| | (0.027) | (0.035) | (0.027) |
| Group size = 5 | -0.265*** | -0.489*** | -0.377*** |
| | (0.028) | (0.041) | (0.029) |
| Quality | 0.726*** | 0.815*** | 0.770*** |
| | (0.011) | (0.014) | (0.012) |
| Distribution Scale | | | -0.101*** |
| | | | (0.015) |
| Intercept | 1.536*** | 1.236*** | 1.436*** |
| | (0.040) | (0.058) | (0.044) |
| Observations | 1,620 | 1,620 | 3,240 |
| $R^2$ | 0.784 | 0.715 | 0.74 |

Robust standard errors clustered on Application in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

for application fixed effects. As the table shows, larger group sizes also lead to more rating accuracy reducing the misalignment between the ex-ante induced quality of the application and the rating, in particular when groups of five applicatins are evaluated jointly. Interestingly, the use of the distribution scale has a very limited effect on rating accuracy.

## Study 3: Noisy but Objective Signals

In a final step we consider a setting in which there is a clear objective standard of performance. To this end we use data from a recent experiment conducted by Kusterer and Sliwka (2024). For this experiment subjects had been hired on Amazon MTurk and were assigned to the role of workers and supervisors. Subjects in the role of workers worked on a real-effort task of entering text from hard-to-read images, similar to "captchas" on 10 consecutive pages decoding 10 images on each page. Subjects in the role of supervisors received a noisy but objective signal of respective employee's performance by learning the percentage of correctly decoded images on a randomly selected subset of the workers' performance outcomes. Supervisors also saw a histogram of all workers' average performances along with the mean and standard deviation. We use the data from two treatments from this study. In both of these treatments supervisors observed one page out of the 10 pages the respective worker had to decode.

Table 3: The Effect of Scale Type and Group Size on the Deviation between Assessed and Induced Quality

| | Deviation between Assessed and Induced Quality | | |
|---|---|---|---|
| | Generic | Distribution | Pooled |
| | (1) | (2) | (3) |
| Group size = 3 | -0.078*** | -0.002 | -0.040 |
| | (0.027) | (0.032) | (0.026) |
| Group size = 5 | -0.235*** | -0.156*** | -0.195*** |
| | (0.027) | (0.033) | (0.024) |
| Distribution Scale | | | -0.013 |
| | | | (0.016) |
| Intercept | 0.748*** | 0.683*** | 0.722*** |
| | (0.027) | (0.028) | (0.026) |
| Observations | 1,620 | 1,620 | 3,240 |
| $R^2$ | 0.025 | 0.011 | 0.016 |

Robust standard errors clustered on application in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

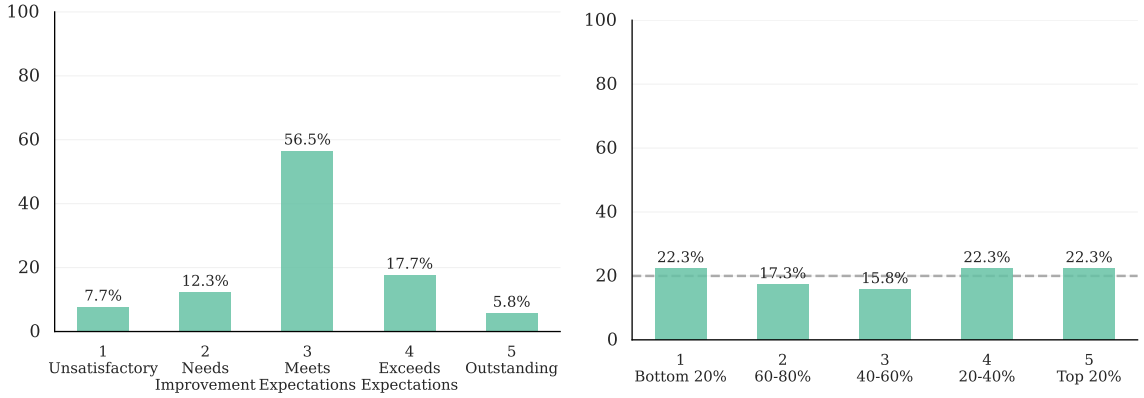
This setting allow us to (i) replicate our previous analyses in a setting where there is an entirely objective performance standard and raters receive noisy performance signals, (ii) compare LLM ratings to a clear Bayesian benchmark of optimal ratings based on rational information processing and, furthermore, (iii) compare LLM ratings to ratings provided by human raters who have exactly the same information structure.

### Comparing Rating Scales

We prompt the LLM to rate the performance of each worker from the experiment based on the information that also had been provided to human supervisors in this experiment, i.e. a description of the task, the number of correctly decoded words on one randomly selected page out of the 10 on which the worker had to work on, as well as information on the distribution of performance outcomes in general.

As a first step we again ask the LLM to apply the rating scales used in the above, i.e. the generic 5-point scale as well as a scale asking to rank the workers in quintiles by their performance. Specifically we used the following text (followed by the respective rating scale as in the above):

> *Your task is to rate the performance of a worker who performed a task on Amazon MTurk. In this task, the worker was shown images and had to enter the text contained in those images accurately under a time constraint. In total, a worker saw 10 pages with 10 images*




(a) Generic Scale

(b) Distribution Scale

Figure 7: Distribution of Ratings by Scale Type

*on each page. Pages had varying time limits between 17 and 25 seconds. Specifically, there were pages with time limits of 17, 19, 21, 23, and 25 seconds, all occurring equally often. That is, some pages were easier to fill in and some were more difficult. The order of the time limits was randomized over the 10 pages.*

*The average performance of 780 workers who completed the Entry Task is 42.5%. This means that on average, they entered the text correctly on 42.5 of the 100 images. The standard deviation of these workers is 15.5.*

*You will see the number of correctly entered images on 1 randomly selected page out of the 10 pages the worker completed whose performance you have to evaluate. You do not know the time limit of this page. The time limit could have been as short as 17 or as long as 25 seconds (or any of the time limits in between these two). You will not learn the worker's performance on the other 9 pages.*

*The worker entered 3 words correctly on the randomly selected page.*

*How would you rate the worker?*

The ratings assigned by the LLM are show in Figure 7. First, note that under bost scales we now observe much weaker rating leniency as compared to our previous results. Apparently the fact that now there is a clear objective performance standard and that the distribution of performance outcomes is known mitigates the tendedency of the LLM to assign lenient ratings. But it is also interesting to note that under the generic scale we still find a rather pronounced central tendency as by far the largest proportion of evaluations (56.5%) is in the middle category. In other words, while it is now easy for the LLM to identify relative performance, it still does not have a clear guideline how good or bad a performance outcome needs to be in order to be evaluated differently to the central "meets expectations" rating. When in doubt, the LLM thus assigns the middle category–just as human raters frequently tend to do under such scales.

But the picture now changes when we move to the distribution scale as shown in the right panel of Figure 7. Here leniency and centrality effects are completely elimiated in the LLM ratings. In this setting the LLM apparently makes use of the information provided and performs rather well in rating performance into the predefined quintiles. Importantly, in contrast to the previous settings this is achieved here despite the fact that the LLM here evaluated individual workers in separate prompts and not in groups, likely because access to the overall performance distribution (mean and standard deviation) facilitates the assignment of these relative ratings.

### Comparison of Human and LLM Ratings

In a final step, we now compare LLM ratings to ratings performed by human raters in this study giving the LLM exactly the same task and information structure as human raters had in the experiment. The evaluators in Kusterer and Sliwka (2024) had beeen tasked with evaluating performance on a scale from zero to one hundred that actually reflects all possible true performance outcomes (as subjects in the role of workers could solve up to 100 decoding tasks). In a next step we thus again use the worker data from the experiment but prompt the LLM with exactly the same intructions as the supervisors in the experiment. That is, we changed the scale in the prompt and added the following sentence that had also been given to the human evalators in the experiment:

> *As a guidance, ratings should reflect the percentage of correctly entered images by the worker across all 10 pages.*

This now allows us to directly compare the rating "behavior" of an LLM to the rating behavior of humans in exactly the same setting. Moreover, in the experiment supervisors had been randomly assigned to a treatment where they were informed that their rating would determine a bonus payment for the worker or to a treatment where no such information was given. In the bonus treatment the following additional information had been provided to human raters (and is now also included in the LLM prompt):

> *The human worker's payment increases in the rating. The worker receives a payment of $1.00 + $2.00 × (your rating)/100.*

Table 4 reports results of regressions of the respective ratings (human in columns (1) & (2) and LLM in columns (3) & (4)) on the signal the raters received (the number of correct solutions on randomly drawn page), a dummy indicating whether the rater was told that the rating determined a bonus, and their interaction. Importantly, human evaluators give significantly more lenient ratings when they know that these ratings determine a bonus payment to the rated agent as columns (1) and (2) show. For a given level of the performance

Table 4: Human vs. LLM Evaluations: Regression Results

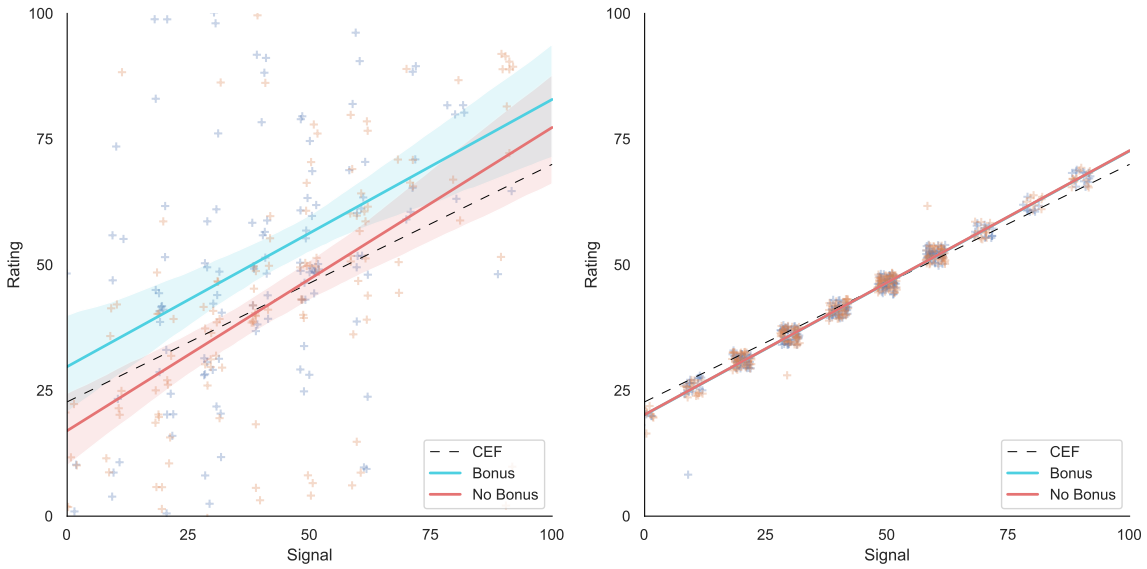
| | Rating by | | | |
|---|---|---|---|---|
| | Human | | LLM | |
| | (1) | (2) | (3) | (4) |
| Signal | 0.573*** | 0.603*** | 0.525*** | 0.525*** |
| | (0.065) | (0.086) | (0.002) | (0.002) |
| Performance pay | 9.732*** | 12.743** | -0.046 | -0.033 |
| | (2.672) | (6.268) | (0.076) | (0.247) |
| Signal × Performance pay | | -0.072 | | -0.000 |
| | | (0.132) | | (0.004) |
| Intercept | 18.298*** | 16.999*** | 20.202*** | 20.195*** |
| | (3.096) | (3.778) | (0.102) | (0.083) |
| Observations | 260 | 260 | 520 | 520 |
| $R^2$ | 0.274 | 0.275 | 0.994 | 0.994 |

Robust standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

signal that the supervisors observed (the share of correct solutions on the one page out of 10 that the supervisors saw) ratings were by *9.73* higher when they determine a bonus to the respective worker. But, as columns (3) and (4) show, the leniency effect of the bonus disappears entirely when ratings are performed by an LLM.

Figure Figure 8 visualizes these results by plotting the average human ratings against the observed performance signals separately by treatment. The dashed black line shows as a rational benchmark the estimated conditional expectations based on the observed signals, estimated through a regression of the true performance on the observed signals.[7] We again see that human ratings are substantially more lenient when there is a bonus. But the slope of the regression line indicates that human evaluators (on average) take the observed performance signals into account in a quite rational way when providing ratings–they just tend to be more generous across the whole performance distribution when they know that the rating affects the well-being of the rated employee. Note, however, that as indicated in the scatter markers the evaluation behavior markedly varies across individual human raters. That is, indidvidual human raters differ substantially in their ratings given the exact same performance signals. But, as shown in panel (b) Figure 8, the LLM acts in a near perfectly rational way when providing ratings: The LLM ratings very closely track the rational benchmark of the conditional expectation of true performance given the observed signals with hardly any noise, irrespective of the consequences for the evaluated employees.

[7]Note that an OLS regression of $y$ on $x$ yields the best linear approximation to the conditional expectation function $E[y|x]$.




(a) Human Ratings

(b) LLM Ratings

Figure 8: Human vs. LLM Evaluations

Finally, we can compare the accuracy of the ratings between the different settings and humans and LLM evaluators. To do this, we compute the rating error as measured by the absolute distance between the rating and the worker's true performance (i.e., the percentage of correctly entered images across all 10 pages). Table 5 shows the results of regression of this rating error on a dummy indicating whether the evaluation was performed by the LLM controlling for the underlying signal noise (i.e,. the absolute deviation between the observed performance signal and the true performance).[8] As both specifications show, the use of an LLM substantially reduces the rating error both for the setting with and without the bonus. Moreover, the gain in rating accuracy from using an LLM is significantly larger when there is a bonus – as here human ratings exhibit the strongest bias in the fist place.

Hence, as our final study documents, LLM can provide highly accurate performance evaluations when there is a clear objective standard of performance and when the distribution of performance outcomes is known. In this setting, LLM ratings closely track rational Bayesian benchmarks and substantially outperform human raters in terms of accuracy, particularly when human evaluations are affected by contextual factors such as incentive structures.

[8]Note that signal noise thus captures what is called the "irreducible error" in machine learning terminology as the noise in the observed performance signal bounds the absolute rating error from below

Table 5: Human vs. LLM Evaluations: Rating quality

| | Rating error | | |
|---|---|---|---|
| | No Bonus | Bonus | All |
| | (1) | (2) | (3) |
| LLM evaluation | -8.867*** | -12.585*** | -8.883*** |
| | (1.238) | (1.462) | (1.237) |
| Signal noise | 0.408*** | 0.355*** | 0.383*** |
| | (0.044) | (0.049) | (0.033) |
| Performance pay | | | 3.631* |
| | | | (1.853) |
| LLM evaluation × Performance pay | | | -3.719* |
| | | | (1.916) |
| Intercept | 12.890*** | 17.213*** | 13.233*** |
| | (1.375) | (1.597) | (1.291) |
| Observations | 390 | 390 | 780 |
| $R^2$ | 0.307 | 0.298 | 0.303 |

Robust standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

## Conclusion

This paper documents how LLMs evaluate performance across settings that systematically vary in the objectivity of underlying performance signals. Our three studies reveal that LLM evaluation behavior depends critically on both signal structure and scale design. When evaluating CEOs without objective benchmarks, LLMs replicate the common centrality and leniency patterns that characterize human performance ratings in organizations. This likely reflects the linguistic priors LLMs inherit from human-generated texts on which they have been trained reflecting the same reluctance to differentiate and to assign too harsh judgements under uncertainty. When assessing job applications with exogenously induced quality differences, LLMs still exhibit systematic upward bias but differentiate more effectively than in purely subjective settings. When, howeverl rating worker performance with clear objective standards and known distributions, LLMs come very close to a fully rational Bayesian benchmark and substantially outperform human evaluators—particularly when human judgments suffer from contextual influences such as incentive-driven leniency (Maas, Rinsum, and Towry 2012; Kusterer and Sliwka 2024).

In the settings without clear objective benchmark, joint evaluations of groups consistently helped to increased rating dispersion, but their effects on leniency diverge by context: In the CEO study, where individual LLM ratings clustered at the midpoint (centrality),

group evaluation increased both dispersion and average ratings, indicating that comparative judgment primarily enabled more generous assessments. In the job application study, where individual ratings concentrated at the upper end (stronger leniency), group evaluation similarly increased dispersion but decreased average ratings. This opposing directional shift reveals that comparative evaluation functions as a context-dependent debiasing mechanism: it does not uniformly push ratings higher or lower, but rather corrects the dominant bias present in individual assessments. When centrality dominates individual ratings, group formats enable differentiation through increased leniency; when leniency dominates, group formats enable differentiation through more critical assessment. The debiasing effectiveness of comparative LLM evaluation appears to depend fundamentally on the initial bias structure of the evaluation context.

These findings carry direct implications for organizations which consider deploying LLMs in performance management. First, signal objectivity emerges as a crucial design parameter: organizations can enhance evaluation quality by providing LLMs with objective performance metrics, distributional information, and clear benchmarks. In the setting we investigated in our third study with noisy but objective performance information LLMs generated much more accurate evaluations than human raters who face stronger cognitive limitations and thus make more mistakes.

Second, comparative evaluation formats—having LLMs evaluate multiple employees simultaneously rather than individually—substantially reduce rating compression, improve differentiation, and importantly also accuracy. This suggests that organizations should structure LLM-assisted evaluations as batch processes rather than isolated judgments. Third, explicit distributional guidance through percentile-anchored scales increases rating dispersion more effectively than generic qualitative labels.

The contrast between LLM and human evaluation behavior in objective settings offers a further useful point of comparison. Where human evaluators show context-dependent leniency—inflating ratings when aware of performance-contingent consequences for employees—LLMs maintained consistent accuracy regardless of the ratings' importance for the workers's well-being. This insensitivity to social considerations implies both opportunities and concerns for practice. Organizations can leverage LLMs when aiming for consistent, unbiased baseline evaluations that remove context-dependent distortions. Human oversight of course can be still valuable for incorporating broader contextual factors, development considerations, and organizational values that extend beyond signal-based accuracy.

Our results demonstrate also that LLM evaluation behavior is not fixed but malleable through prompt design. Organizations adopting these systems should view evaluation quality as an engineering challenge: by carefully structuring the information environment, rating scales, and comparison sets, practitioners can systematically improve LLM evaluation

performance (Bol, Brown, and LaViers 2025). The substantial quality gains we document from objective signals, comparative formats, and distributional anchoring suggest concrete pathways for organizations to enhance AI-assisted performance management. As LLMs become more prevalent in organizational decision-making, understanding how to structure evaluation tasks in order to ensure these systems serve organizational goals effectively becomes crucial.

## References


Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci. 2024. "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech."

Baker, George, Michael Gibbs, and Bengt Holmstrom. 1994. "The Internal Economics of the Firm: Evidence from Personnel Data." *The Quarterly Journal of Economics* 109 (4): 881–919.

Becker, Geraldine A, and Charles E Miller. 2002. "Examining Contrast Effects in Performance Appraisals: Using Appropriate Controls and Assessing Accuracy." *The Journal of Psychology* 136 (6): 667–83.

Berger, Johannes, Christine Harbring, and Dirk Sliwka. 2013. "Performance Appraisals and the Impact of Forced Distribution—an Experimental Investigation." *Management Science* 59 (1): 54–68.

Bol, Jasmijn C. 2008. "Subjectivity in Compensation Contracting." *Journal of Accounting Literature* 27: 1–24.

Bol, Jasmijn C., Conor Brown, and Lisa LaViers. 2025. "Context and Design Matter: Employee Acceptance of Using Artificial Intelligence to Conduct Performance Evaluations." Available at SSRN: https://ssrn.com/abstract=5112223.

Bond, Brittany M. 2025. "Cut to the Curve: Underrecognition and Talent Loss from Forced Ranking in a Multinational Firm." *Management Science*.

Bretz Jr, Robert D, George T Milkovich, and Walter Read. 1989. "Comparing the Performance Appraisal Practices in Large Firms with the Directions in Research Literature: Learning More and More about Less and Less." *Cornell University, School of Industrial and Labor Relations Working Paper* 89-17.

———. 1992. "The Current State of Performance Appraisal Research and Practice: Concerns, Directions, and Implications." *Journal of Management* 18 (2): 321–52.

Cardinaels, Eddy, and Christoph Feichter. 2021. "Forced Rating Systems from Employee and Supervisor Perspectives." *Journal of Accounting Research* 59 (5): 1573–1607.

Dargnies, Marie-Pierre, Rustamdjan Hakimov, and Dorothea Kübler. 2024. "Behavioral Measures Improve AI Hiring: A Field Experiment." In *Proceedings of the 25th ACM Conference on Economics and Computation*, 831–32.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey. 2018. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them." *Management Science* 64 (3): 1155–70.

Dohmen, Thomas J, Ben Kriechel, and Gerard A Pfann. 2004. "Monkey Bars and Ladders: The Importance of Lateral and Vertical Job Mobility in Internal Labor Market Careers." *Journal of Population Economics* 17 (2): 193–228.

Flabbi, Luca, and Andrea Ichino. 2001. "Productivity, Seniority and Wages: New Evidence

from Personnel Data." *Labour Economics* 8 (3): 359–87.

Frederiksen, Anders. 2013. "Incentives and Earnings Growth." *Journal of Economic Behavior & Organization* 85: 97–107.

Frederiksen, Anders, Fabian Lange, and Ben Kriechel. 2017. "Subjective Performance Evaluations and Employee Careers." *Journal of Economic Behavior & Organization* 134: 408–29.

Frederiksen, Anders, and Előd Takáts. 2011. "Promotions, Dismissals, and Employee Selection: Theory and Evidence." *The Journal of Law, Economics, & Organization* 27 (1): 159–79.

Goffin, Richard D, and James M Olson. 2011. "Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others." *Perspectives on Psychological Science* 6 (1): 48–60.

Golman, Russell, and Sudeep Bhatia. 2012. "Performance Evaluation Inflation and Compression." *Accounting, Organizations and Society* 37 (8): 534–43.

Heneman, Robert L. 1986. "The Relationship Between Supervisory Ratings and Results-Oriented Measures of Performance: A Meta-Analysis." *Personnel Psychology* 39 (4): 811–26.

Horton, John J. 2017. "The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment." *Journal of Labor Economics* 35 (2): 345–85.

Jabarian, Brian, and Luca Henkel. 2025. "Voice AI in Firms: A Natural Field Experiment on Automated Job Interviews." *Firms: A Natural Field Experiment on Automated Job Interviews (August 18, 2025)*.

Jawahar, I. M., and Charles R. Williams. 1997. "Where All the Children Are Above Average: The Performance Appraisal Purpose Effect." *Personnel Psychology* 50 (4): 905–25.

Kantor, Jodi, and Arya Sundaram. 2022. "The Rise of the Worker Productivity Score." *The New York Times*. https://www.nytimes.com/interactive/2022/08/14/business/worker-productivity-tracking.html.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1): 237–93.

Kusterer, David J, and Dirk Sliwka. 2024. "Social Preferences and the Informativeness of Subjective Performance Evaluations." *Management Science*.

Landy, Frank J., and James L. Farr. 1980. "Performance Rating." *Psychological Bulletin* 87 (1): 72–107.

Li, Danielle, Lindsey Raymond, and Peter Bergman. 2025. "Hiring as Exploration." *Review of Economic Studies*, rdaf040.

Ludwig, Jens, and Sendhil Mullainathan. 2024. "Machine Learning as a Tool for Hypothesis Generation." *The Quarterly Journal of Economics* 139 (2): 751–827.

Maas, Victor S., Marcel van Rinsum, and Kristy L. Towry. 2012. "In Search of Informed Discretion: An Experimental Investigation of Fairness and Trust Reciprocity." *The Accounting Review* 87 (2): 617–44.

Manthei, Kathrin, and Dirk Sliwka. 2019. "Multitasking and Subjective Performance Evaluations: Theory and Evidence from a Field Experiment in a Bank." *Management Science* 65 (12): 5861–83.

Moers, Frank. 2005. "Discretion and Bias in Performance Evaluation: The Impact of Diversity and Subjectivity." *Accounting, Organizations and Society* 30 (1): 67–80.

Mullainathan, Sendhil. 2025. "Economics in the Age of Algorithms." In *AEA Papers and Proceedings*, 115:1–23. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Murphy, K. R., and J. N. Cleveland. 1995. *Understanding Performance Appraisal*. Thousand Oaks: Sage.

Ockenfels, Axel, Dirk Sliwka, and Peter Werner. 2015. "Bonus Payments and Reference Point Violations." *Management Science* 61 (7): 1496–1513.

———. 2025. "Multirater Performance Evaluations and Incentives." *Journal of Labor Economics* 43 (4): 985–1004.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37: 7–63.

Prendergast, Canice, and Robert Topel. 1996. "Favoritism in Organizations." *Journal of Political Economy* 104 (5): 958–78.

Tambe, Prasanna, Peter Cappelli, and Valery Yakubovich. 2019. "Artificial Intelligence in Human Resources Management: Challenges and a Path Forward." *California Management Review* 61 (4): 15–42.

Wagner, Stephen H, and Richard D Goffin. 1997. "Differences in Accuracy of Absolute and Comparative Performance Appraisal Methods." *Organizational Behavior and Human Decision Processes* 70 (2): 95–103.

# Appendix

## Study 2: Prompt Template

Here is the prompt template we used to generate job application texts of varying quality levels:

> *Write a application text for a recent university graduate applying for a {job_title} position. The text should clearly reflect a quality level of {quality_text} in this target group. Write the application as a single paragraph with no line breaks. Do not include greeting, closing, or name — provide only the main text, and keep it under 200 words.*

Quality Text:

> *1 = The application is among the bottom **20%** of applications.*
>
> *2 = The application is among the bottom **40%** of applications but not the bottom **20%** (4th quintile, , second-worst quality tier).*
>
> *3 = The application is among the middle **20%** (3rd quintile) of applications.*
>
> *4 = The application is among the top **40%** of applications but not the top **20%** (2nd quintile, , second-worst quality tier).*
>
> *5 = The application is among the top **20%** of applications.*

To illustrate the quality levels used in Study 2, we provide examples of generated application texts. These were created by prompting GPT-5-mini to write applications at specified quality levels, then evaluated by a separate LLM instance without knowledge of the induced quality.

### Example 1: Very Bad Quality (Bottom 20%)

**Job Title:** Junior Business Analyst

**Application Text:**

> *I recently graduated from university with a business degree and I am applying for the Junior Business Analyst position because I need a start in this field and I think this job could be good for me, I don't have much formal work experience in analysis but I did group projects and assignments where I used Excel a bit and made some reports, I am familiar with basic statistics from coursework and I can learn more software quickly, I am punctual and willing to work long hours, I have good communication in English and some presentation practice, I am open to training and flexible about tasks and salary, I hope you will consider me even though I am new to the role because I want to grow and probably contribute in small ways while I learn, thank you for reading this application.*

**Example 2: Very Good Quality (Top 20%)**

**Job Title:** Junior Business Analyst

**Application Text:**

*With a Bachelor's degree in Business Analytics and graduation in the top 10% of my cohort, I bring strong analytical ability, practical technical skills, and a results-focused mindset to the Junior Business Analyst role. In internships and capstone projects I developed SQL-based data pipelines, automated reporting in Excel and Power BI, and used Python for regression and segmentation analyses that uncovered process improvements and improved forecast accuracy by up to 15%. I pair quantitative skills (SQL, Excel, Power BI, Python, Tableau) with business coursework in finance, operations and strategy, and have presented findings and recommendations to stakeholders across functions. I excel at defining root causes, turning ambiguous questions into testable hypotheses, and delivering prioritized, actionable insights that support decision-making. Fast-learning, collaborative, and detail-oriented, I am motivated to contribute analytical rigor and efficient reporting to drive measurable improvements at your organization.*

## Additional Figures

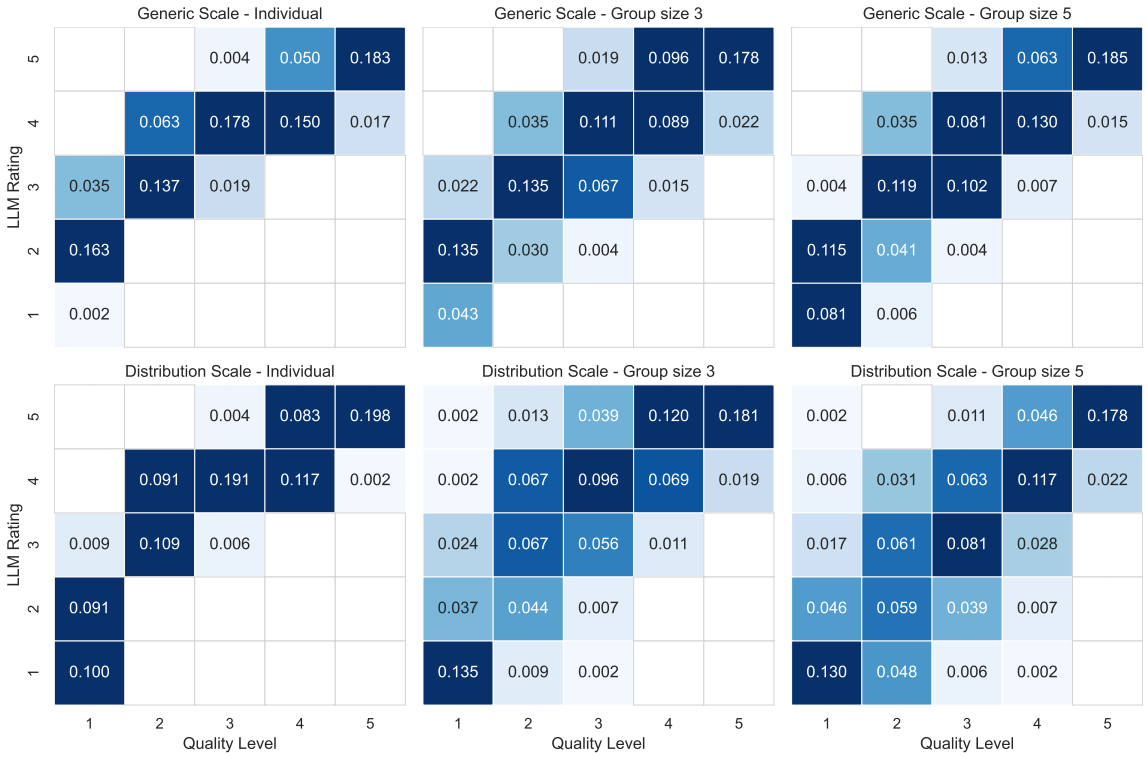



Figure A1: Confusion Matrices: LLM Ratings vs induced Quality Levels