PYX Labs.ai

PYX-Voice

Sarah Foster **Megan Steckler** **Sumit Gupta** **Andy Horng**
Lauren Beechly **Ethan Burris** **Joseph Freed**

Abstract

*We introduce **PYX-Voice**, a benchmark for assessing employee listening analysis and interpretation. **PYX-Voice** requires LLMs to navigate realistic workplace tasks with datasets and files. We test 7 LLMs for the leaderboard using Inspect. Gemini-3.5-flash achieves the highest score of 76%, followed by gpt-5-2025-08-07 (71%), and grok/grok-4 (70%). More information about the benchmark is accessible via www.pyxlabs.ai.*

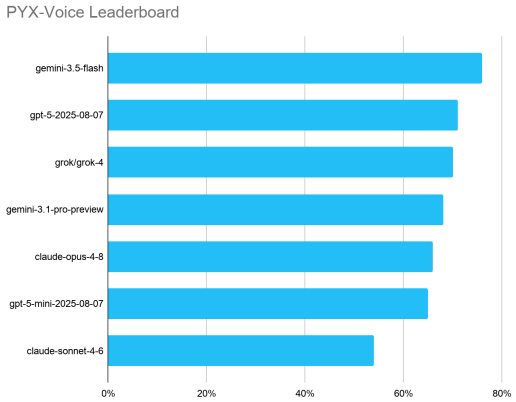


***Figure 1.** Model performance across the entire benchmark. Values reflect the share of all 84 tasks passed.*

1. Introduction

The use of large language models (LLMs) in workplace settings has accelerated rapidly, driven by competitive pressure to implement emerging technologies (Gans, 2023). Many organizations are investing significant resources in AI adoption (Singla et al., 2025; Cramer, 2026). However, adoption is outpacing evaluation: the return on AI investment is not yet fully understood, necessitating rigorous exploration of LLM performance on domain-specific tasks (Olson, 2026). The Human Resources (HR) field is no exception to this widespread trend (SHRM, 2026). Additionally, LLM applications in the HR domain carry distinct risks that the broader benchmarking literature has not yet addressed.

HR practitioners work with sensitive data about real people and use it to design interventions at the individual, team, and organizational levels. One common practice in HR teams is to measure and take action related to employee experience: this work often involves a robust data collection process in which HR teams gather employee sentiment data, and then use those large datasets to inform decisions, interventions, and actions. These employee experience interventions are meaningful: research consistently links employee experience to individual life satisfaction and to organizational performance (Kossyva et al., 2023; Bateman, 2026).

1.1 Risks of AI Application

In the HR domain, the stakes of AI errors are concrete and consequential. The cost of inaccurate analysis ranges from a misdirected team meeting to an executive decision that undermines business performance, results in significant financial loss, or inadvertently discriminates against protected employees.

On the analysis side, LLMs are known to produce calculation errors and hallucinations: failure modes with direct consequences when outputs drive decisions about real employees. What remains unclear is the frequency and severity of these errors in employee experience specific tasks, which models perform best, and what safeguards effectively reduce risk.

On the reporting and interpretation side, LLMs tend toward verbosity ill-suited to executive-level communication, which demands precise data storytelling that integrates quantitative findings, qualitative themes, business priorities, and industry context into a coherent, actionable narrative. There is also a deeper risk: employee listening recommendations should be grounded in the scientific literature on human and organizational behavior: the domain of industrial-organizational (I/O) psychology. LLMs risk substituting popular management trends or low-quality sources for evidence-based guidance, with consequences that many practitioners are not equipped to detect.

1.2 Opportunities of AI Application

HR practitioners working in the employee experience domain face a well-documented insights-action gap: the persistent lag between collecting employee listening data and translating it into organizational action (Burris et al., 2024; Perceptyx, 2023). Research on employee voice shows clear benefits for organizations, leaders, and employees. Across industries, voice is linked to stronger organizational performance: a 2014 study of 38 hospitals found that hospitals that effectively encouraged feedback from customer relations employees received service-performance ratings 27% higher from CEOs and 41% higher from vice presidents (Lam & Mayer, 2014). Leaders also benefit when employee input reaches them directly; in financial service organizations, units where voice flowed to leaders were 16% more financially and operationally effective than units where ideas circulated around leaders but did not reach them (Detert et al., 2013). Employees benefit as well: when organizations give people fair, consistent ways to participate in decisions, they report more positive feelings about their jobs, greater intrinsic motivation, and lower attrition—about 50% lower in units where managers effectively support voice and encourage new ideas (Zapata-Phelan et al., 2009; McClean et al., 2013).

Yet creating a culture of voice remains difficult. Many employees are hesitant to participate in employee surveys because of two primary reasons: 1) a sense of fear of speaking the truth and relaying critical feedback, and 2) a sense of futility that participating in the employee listening process will not produce substantive change. Managers can compound the problem by discouraging or ignoring the input they say they need, especially when it feels threatening or challenges established routines. For example, a previous study found that managers were 69% less likely to endorse subordinate ideas for implementation that challenged the status quo (Burris, 2012).

The key challenge then for business leaders is to manage the employee listening experience in such a way that extracts useful insights for executives, protects the confidentiality and safety of individual employees, and propels the organization to take action to address issues that employees have raised.

The process is labor-intensive, involving extensive analysis, cross-functional alignment meetings, and iterative reporting cycles, while employees grow skeptical that their feedback produces any change at all. AI offers a credible path to closing this gap: accelerating analysis, surfacing patterns across large datasets, and

enabling faster, more responsive action-taking grounded in both organizational context and evidence-based practice.

1.3 Evaluating AI Performance

Despite these risks and opportunities, HR practitioners currently lack clear standards for evaluating AI performance in this domain: what "good" looks like, which failure modes matter most, and when AI assistance is appropriate versus insufficient. This is consistent with other domains: while benchmarking efforts are becoming more commonplace to define and promote responsible AI, those efforts are not keeping pace with the capability advancements (Sajadieh et al., 2026). This gap produces uneven adoption, with some practitioners over-relying on AI outputs without accounting for the inherent risks and others avoiding the technology entirely, leaving significant productivity and quality gains unrealized.

What is needed is a domain-specific benchmark developed by field experts and rigorous enough to meet the standards of the AI research community. As a leading authority in employee experience listening, PYX Labs is uniquely positioned to fill this gap. With a proprietary database representing millions of employee responses and deep in-house expertise spanning I/O psychology, people analytics, and HR consulting, PYX Labs enlisted in-house subject matter experts (SMEs) to develop PYX-Voice: the first benchmark for assessing LLM performance on employee listening analysis and interpretation tasks.

To construct PYX-Voice, we created three "worlds": realistic datasets representing 17,500 employees across three hypothetical organizations. SMEs with deep EX listening expertise designed realistic tasks of varying complexity and difficulty, drawn from actual practitioner workflows. SMEs then completed the tasks themselves and provided evaluation criteria, establishing human expert performance baselines. AI tools were then used to help formalize evaluation criteria, which SMEs reviewed and approved to ensure both rigor and practical validity.

The result is a collection of 84 unique tasks across three datasets, with two task types assessing seven distinct LLM capabilities across three difficulty levels. The benchmark encompasses 208 discrete evaluation criteria. More information about PYX-Voice is available at www.pyxlabs.ai.

2. Benchmark Design and Dataset

2.1 Dataset Construction

PYX-Voice is built on three datasets representing the employee experience survey output of 17,500 real employees (n = 3,000, n = 8,500, n = 6,000). Two of the datasets included longitudinal data with two data collection rounds each (i.e., dataset 2 from 2021 and 2023, dataset 3 from 2025 and 2026). Each dataset contains 35 - 76 survey items per data collection period, including demographics along with a combination of Likert-type quantitative items and open-ended qualitative responses, which is standard industry practice.

Each dataset represents one hypothetical organization, for which in-house SMEs developed a unique contextual brief containing realistic business priorities, organizational characteristics, and key analysis goals, ensuring that tasks were grounded in a coherent organizational scenario rather than isolated data points. Each dataset was matched with external benchmark comparisons, which were used as additional context for task completion.

PYX Labs utilizes anonymized data from customers in low-risk jurisdictions who have opted into benchmarking. We aggregate and de-identify data across multiple sources, extracting random samples to ensure that no single customer's information is identifiable or reconstructible. Under continuous legal review, we apply further data transformations to guarantee that our training sets remain entirely

anonymous and compliant with global privacy standards.

2.2 Task Design

Task development began with a formal job analysis of the employee experience consultant role: the human practitioner whose work PYX-Voice tasks are designed to replicate. This analysis drew on job description review, consulting deliverable review, SME interviews, and O*NET occupational data to identify the core task requirements, skill demands, and performance expectations of the role. Grounding task design in validated occupational science rather than ad hoc expert judgment ensures that PYX-Voice assesses LLM performance on tasks that reflect real practitioner work.

From this analysis, SMEs created 84 realistic and challenging tasks, spread across three hypothetical organizations, that align with **key components of the employee listening process**:

- **Data analysis:** Performing statistical analyses to summarize results, extracting themes from qualitative data to provide color and depth to the analyses
- **Data interpretation:** Creating narratives from the data analysis and results, crafting executive summaries

In other words, these tasks involve not only extracting the information from datasets, but also judgment and nuance in extracting the right information and interpreting it the right way. These tasks also span the following **LLM capability dimensions**:

- **Calculation:** Arithmetic operations including percentages, demographic breakdowns, and/or statistical analyses
- **Retrieval:** Extracting verbatim responses from qualitative datasets
- **Summarization:** Condensing open-ended comments into coherent themes
- **Synthesis:** Integrating quantitative and/or qualitative data with domain knowledge to produce action-oriented outputs
- **Multi-step instruction following:** Executing two- or three-part task sequences
- **Domain expertise:** Applying terminology and concepts from I/O psychology and behavioral science
- **Data integrity and anomaly detection:** Identifying errors, inconsistencies, or irregularities in datasets

Tasks are distributed across three difficulty levels (low, medium, and high), and two task types: Verifiable Response and Summary. These **task types** are conceptualized as:

- **Verifiable Response:** Typically based on quantifiable metrics, these tasks have a binary correct or incorrect response.
- **Summary:** Typically based on open-ended employee feedback, these tasks involve summarizing data to arrive at thematic conclusions.

In a real-world context, it is estimated that completing this collection of 84 tasks would require approximately 20 hours of SME time – a figure that underscores both the depth of the benchmark and the efficiency gains AI could provide if it performs at an acceptable standard.

2.3 Rubrics and Gold Standard Outputs

For each task, SMEs completed the task themselves, producing a gold standard response alongside specific performance criteria. This human expert baseline is the standard against which all LLM outputs are evaluated.

For verifiable response tasks, evaluation criteria are objective: the output is either correct or it is not (e.g., *What item in the current survey results most exceeds the industry benchmark, and by how much?*). Summary tasks are more subjective by nature (e.g., *Provide a 2-3 sentence summary for the comment item "What additional feedback do you have?" around the themes of senior leadership.*). For summary tasks, SMEs provided example outputs alongside 3–5 task-level criteria organized across two **evaluation dimensions**:

Data Integrity and Accuracy

- Does not hallucinate
- Correctly identifies errors or anomalies in the data
- Accurately identifies relevant themes

Communication of Insights

- Produces coherent, well-structured narratives
- Synthesize recommendations aligned to stakeholder priorities/strategy (internal context), external context, and data insights
- Provides actionable guidance
- Adopts appropriate executive tone and register

A preliminary third evaluation dimension was also developed, but not yet validated for this benchmark. Future research will include criteria aligned to ***Technical and Domain Expertise*** (e.g., demonstrates alignment with behavioral science principles, exhibits appropriate bias and sensitivity awareness, and contextualizes findings within relevant industry conditions).

During task and criteria development, all content underwent a structured peer review and calibration process. Each task and criteria pairing was reviewed and approved by a minimum of two SMEs before inclusion in the benchmark, ensuring consistency and quality across the full task set.

2.5 AI-Enabled Calibration & Refinement

Given the number of contributors in this work, AI tools were used to identify inconsistencies and clarify best practices for prompts and criteria in early stages of task and criteria development.

In addition, AI tools were used to refine and standardize final criteria for more objective, repeatable measurement and scoring. Following the AI-assisted standardization, SMEs underwent two rounds of human grading to validate the benchmark.

For the first round, the most experienced SME reviewed a stratified random sample of final tasks and criteria, alongside LLM task output from multiple LLMs (n=150). In that review, they: (1) Confirmed final criteria are relevant and accurate. In the rare instances where the SME disagreed with the final task or criteria content, they provided detailed explanations which were used to refine the criteria further; they also (2) graded sample LLM task output against the final criteria, as if they were assessing a consultant's work. In the instances where human ratings significantly differed from AI ratings, the SME provided detailed feedback and rationale which were used to refine the scoring algorithm. Inter-rater reliability analyses on the Verifiable Response grading revealed human-to-LLM Judge agreement at $\kappa \approx 0.66$.

In the second round, three SMEs blindly reviewed a stratified set of Summary outputs (n = 48), providing pass/fail scores for each along with commentary for each fail. This round of review enabled us to fine-tune the Themes judge, which focuses on the accuracy of themes selected for a given task. Inter-rater reliability analyses on the Themes grading revealed human-to-human agreement at $\kappa \approx .70$ and human-to-LLM Judge agreement at $\kappa \approx 0.51$.

These rounds of SME review and grading also provided preliminary data for criteria related to executive readiness, technical expertise, and actionability, which will be further explored in future research.

2.6 Contributors

Subject matter experts (SMEs) were selected based on their analytical expertise, job performance, and years of experience in analyzing and interpreting employee experience data.

A total of eight SMEs contributed to the work, four with masters-level education in relevant fields (e.g., I/O psychology, MBA), three with PhD-level education in I/O psychology. Experts' mean relevant experience is 14 years. Experts worked as both contributors and reviewers; creating tasks, pulling data, auditing quality, grading output, and providing and checking gold outputs and rubrics.

3. Evaluation

3.1 LLM-as-judge Protocol

PYX-Voice uses a rubric-based LLM-as-judge methodology to evaluate both Verifiable Response and Summary task types. To evaluate the model output for a certain task, a single judge model is chosen from a different vendor family than the model under evaluation and used across all relevant judges for that task type. For example, an OpenAI model response is never graded by an OpenAI judge model.

For Verifiable Response tasks, a single Verifiable Response Judge is used to determine the accuracy of each model response. For the more open-ended Summary tasks, multiple judges are used for evaluation, each with distinct rubrics aligned to a specific evaluation dimension. This structured approach allows us to isolate specific failure modes (i.e. thematic errors, model hallucinations, and low quality of communication) and transparently report model strengths and weaknesses. A Summary task only passes if it clears all judges.

Scores reported throughout are pass rates — the share of tasks a model passed — aggregated overall and by task type, capability, analysis type, and employee-experience area.

3.2 Verifiable Response Judge (Verifiable Response Tasks)

Verifiable Response tasks, which have an objectively correct answer, are evaluated solely by the Verifiable Response judge. The judge compares the model's response to the SME gold answer and returns a binary correct/incorrect verdict. Internally, the judge utilizes a rubric (static across all tasks) which checks that each required element of the SME gold answer is provided, and that there are no claims in the model output which contradict any aspect of the SME gold answer.

3.3 Themes Judge (Summary Tasks)

A Themes judge is used to evaluate model outputs for Summary tasks, with a custom task-specific rubric defined by the SME. The rubric for each task covers key criteria around data accuracy (that the model correctly identifies errors/anomalies), thematic accuracy (that the model correctly communicates key summary points), and actionability (that the model correctly synthesizes actionable recommendations aligned to stakeholder priorities).

The Themes judge prompt and task-level criteria were fine-tuned against human evaluations provided by three SMEs, working independently across all Summary tasks. We tuned the judge until its pass/fail decisions aligned with expert judgment.

3.4 Hallucination Judge (Summary Tasks)

A Hallucination judge is used to evaluate Summary tasks, to verify if model outputs stay faithful to the underlying survey data with no fabricated numbers or quotes. The judge is provided access to the underlying data through tool-calling, and utilizes a contraction rubric (static across all tasks) which recomputes high-risk aspects of the model response (i.e. favorability scores, benchmark deltas, and quoted comments) and checks if the recomputed values materially contradict the model response.

3.5 Foundational Progress: Executive Readiness Judge (Summary Tasks)

In this iteration of the benchmark, we began preparing an Executive Readiness Judge. However, it was omitted from this leaderboard scoring to undergo additional updates and validation. Future versions of this benchmark will use this judge to evaluate Summary tasks on quality of communication (as opposed to content). The judge will also use a rubric (static across all tasks) to determine if a model's response to a Summary task is clearly formatted (i.e. headline conclusion appears at the beginning of the response and summary points are demarcated by bullet points), has no extraneous information (i.e. no methodology preamble before the conclusion and uses standard EX terminology), and is actionable (i.e. recommendations have clear owners, time horizons, and definitions of success).

4. Results

4.1 Overview

Scores are reflected in percentage form, representing the share of tasks passed out of the whole. Across the full benchmark (n = 84 tasks; Figure 1), we found gemini-3.5-flash to be the highest overall performer (76%), with gpt-5-2025-08-07 (71%) and grok/grok-4 (70%) following. The lowest-scoring model on the full benchmark was claude-sonnet-4-6 (54%).

4.2 Task Type

Narrowing the focus to only the Verifiable Response tasks (n = 55; Figure 2), the majority of models performed within a narrow range (76% - 82%), with claude-sonnet-4-6 trailing behind in last place (64%). Gemini-3.5-flash, gpt-5-2025-08-07, and grok/grok-4 were tied as the highest overall performers (82%). Claude-opus-4-8 was close behind (80%).

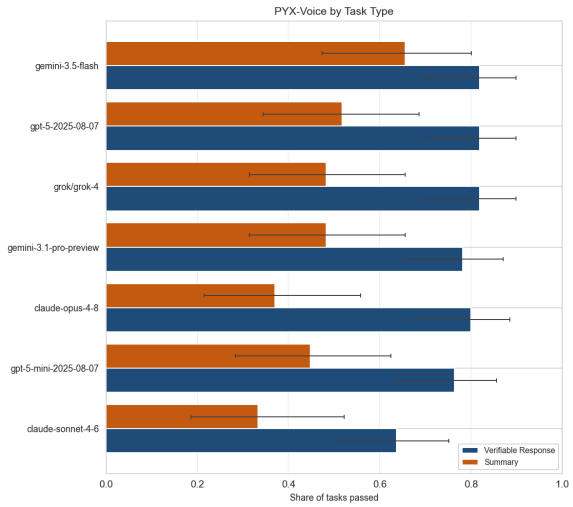


Figure 2. *Model performance by task type. Values reflect the share of tasks passed per task type. Verifiable Response (n = 55) tasks are typically based on quantifiable metrics, with a binary correct or incorrect response. Summary (n = 29) tasks are typically based on open-ended employee feedback, involving summarizing data to arrive at thematic conclusions.*

For the Summary tasks (n = 29; Figure 2), model performance was more wide-ranging (33% to 66%). Overall, models did not perform as well on these tasks as they did for the Verifiable Response tasks. Gemini-3.5-flash was the top performer (66%), with gpt-5-2025-08-07 (52%), gemini-3.1-pro-preview (48%), and grok/grok-4 (48%) following. The lowest-scoring model on the Summary tasks was claude-sonnet-4-6 (33%).

4.3 LLM Capability

Regarding the LLM capabilities assessed, models generally scored highest on Calculation (67% - 97%), and lowest on Synthesis (14% - 57%; Figure 3). The capability with the widest distribution was Retrieval, with claude-sonnet-4-6 (25%) at the low end and two models tied at 100%: gemini-3.5-flash and gemini-3.1-pro-preview. Data integrity/anomaly detection also had a large distribution of performance, with multiple models on the low end (50%: gpt-5-2025-08-07, grok/grok-4, and gpt-5-mini-2025-08-07), and gemini-3.5-flash on the high end (100%). The capability with the most narrow distribution was Multi-Step

Instruction Following (54% - 77%). For a deeper understanding of LLM capabilities, tasks requiring analytical processes were tagged with the type of analysis used (Figure 4). Models generally scored highest on tasks requiring benchmark comparisons (comparing values to external reference points that were provided to LLM), and lowest on tasks requiring trend comparisons (comparing Year 1 data to Year 2 data). However, given the low representation of each of these analytical types, these results are best interpreted directionally. Narrowing the focus to analytical techniques that were represented in five or more tasks, models generally scored highest on tasks requiring descriptive analyses, and lowest on tasks requiring response-based filtering.

4.4 Employee Experience

Each task was aligned with up to two employee experience areas based on the Perceptyx People Insights Model (Figures 5a, 5b, 5c). Models generally scored highest on tasks related to DEIB, Performance Enablement, and Workplace Wellbeing. In contrast, models scored lowest on tasks related to Future/Vision, Change & Innovation, Growth & Development, and Engagement. It is worth noting that two factors (Growth & Development and Teamwork & Collaboration) were represented in fewer than 5 tasks - thus, results for these factors should be interpreted directionally.

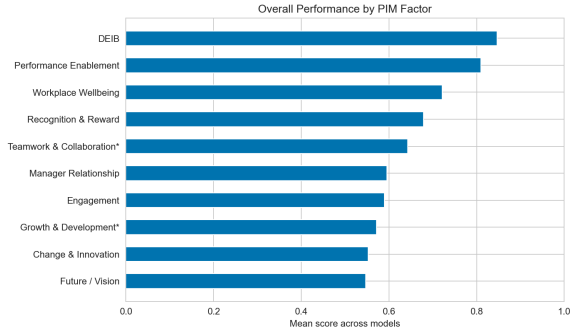


Figure 5a. *Percentage values reflect the mean scores across models for each employee experience factor in the Perceptyx People Insights Model. *Indicates factors with fewer than 5 tasks. Results should be interpreted directionally.*

4.5 Critical Failures

Critical failures were rare but present (Figure A1). Most models had two overclaim failures across the 84 tasks, with the exception of grok/grok-4 (which had one). Claude-sonnet-4-6 fabricated one statistic.

4.6 Cost & Latency

Models varied in terms of the median duration (24 - 132 s), tokens in (406K - 4.9 M), and tokens out (111K - 1.9 M; Figure A2). Claude-sonnet-4-6 was a stark outlier in terms of both latency (132 s) and tokens out (621K), roughly five times slower than the fastest models and three times more output tokens than the next model. Gemini-3.5-flash, in contrast, queried the data heavily (4.9 M tokens in, ~3x the next model), with efficient (24 s) and concise output (238K tokens out).

5. Discussion

The PYX-Voice benchmark is the first benchmark for assessing LLM performance on employee listening analysis and interpretation tasks. Thus, the findings from this research are preliminary and open further avenues for discussion and exploration.

Broadly, gemini-3.5-flash consistently performed best and claude-sonnet-4-6 performed worst regardless of task type. Models generally performed better on Verifiable Response tasks than Summary tasks. Verifiable Response tasks are more objective, with clear right-or-wrong answers. Summary tasks, in contrast, are more subjective, giving greater license for each model to approach the task in its own unique manner. This subjectivity accounts for the wider range and variability of model performance for these tasks.

From an HR practitioner perspective, it was unexpected that the LLMs performed so well on Verifiable Response tasks, which required calculations and statistical analyses. While critical errors did occur, they were rare. The

rarity surprised our SMEs, who expressed assumptions shared within the HR community that “AI doesn’t do math well.” Nevertheless, even with the generally strong performance, there remains a low tolerance level for errors in this work domain, indicating that HR practitioners may need more assurances and higher performance to trust models to statistically analyze quantitative data.

The bigger opportunity for saving practitioner time, however, lies in the Summary tasks – the subjective, complex work of distilling key themes and insights from datasets. LLMs performed reasonably well at identifying relevant themes, but performance varied widely across tasks, suggesting AI is not yet ready to take this work off practitioners’ hands. That gap matters: HR practitioners face real pressure to adopt AI for these time-intensive tasks, and the potential time savings are substantial. But our findings suggest that leveraging AI for this kind of nuanced, high-impact work risks unpredictable performance - and with it, poor organizational decisions. Preliminary anecdotal evidence from our work on executive readiness and action-taking suggests this risk grows as more nuance and practical considerations come into play. HR practitioners using AI to derive meaning from complex data should therefore keep a human in the loop at critical decision points.

Interestingly, our results suggest that heavier reasoning capability didn't translate into better performance on these tasks. Gemini-3.5-flash, a fast, cost-efficient model, was the top overall performer (76%) and led the subjective Summary tasks (66%), while Claude Opus 4.8, generally positioned for complex reasoning, placed mid-pack (5th, 66%) and was notably weaker on Summary (37%). It is also worth noting that reasoning-effort settings differed by vendor (Anthropic/OpenAI ran at medium, Gemini/Grok at low); even with this variability, Gemini was the highest performer at low effort.

This same pattern holds at the capability level: across all models, Synthesis – the work of pulling disparate data elements together for interpretation – was the lowest-scoring capability we measured. Together, these findings suggest that raw reasoning power isn't the bottleneck; the harder challenge is synthesis itself, a distinct capability that current models haven't mastered regardless of how much computation they're given. PYX Labs plans to study this gap directly, with future research focused on frontier models' capacity to synthesize data and recommend actions from survey results.

That capability gap also shows up when we break results down by topic. We applied the Perceptyx People Insights Model framework to each task to explore trends by employee experience factor. Results from this analysis are best interpreted directionally, as each factor was not equally represented across tasks. Still, consistent with our broader findings, LLMs performed most poorly with topics that are complex and nuanced (Future/Vision, Change & Innovation, Growth & Development, Engagement). In contrast, models performed best on tasks related to topics that are more clear, both in how survey respondents articulate their concerns and in how they can be thematically categorized and addressed (DEIB, Performance Enablement, Workplace Wellbeing).

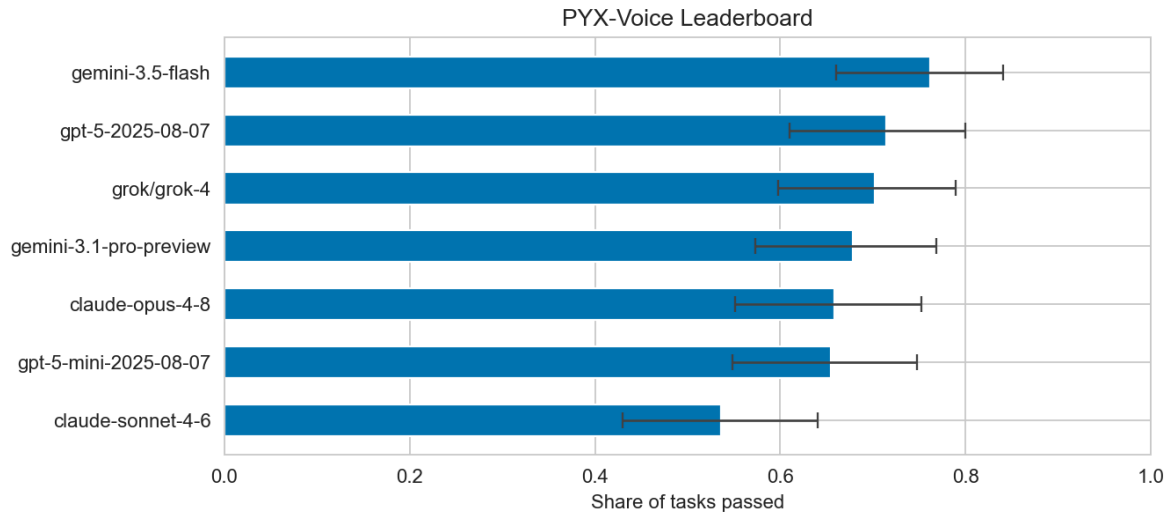
For example, employee feedback related to Performance Enablement typically has very clear, consistent terminology (e.g., goals, resources, tools, success metrics) that easily maps to this category. In contrast, employee feedback related to Change & Innovation can be broad, complex, and nuanced, reflecting not only an employee’s personal experience of a particular organizational change, but also their unique human reaction to that change. A seasoned human consultant would be able to interpret and categorize feedback relevant to this topic, whereas LLMs may need more guidance and context to do so.

This benchmark is the first of its kind in this content domain, and its results are best read as

a starting point rather than a verdict. Frontier models can already handle a meaningful share of employee listening analysis – particularly the objective, verifiable work – but the more valuable work of interpretation and synthesis still exceeds what these models can reliably deliver. Closing that gap is the central challenge for AI in HR advisory work, and it's where PYX Labs will focus next: expanding our Summary task criteria beyond theme identification into more subjective, high-stakes territory like executive readiness, technical expertise, and actionability. Until then, the organizations that get the most value from AI in this space will be the ones that know where to still trust their practitioners.

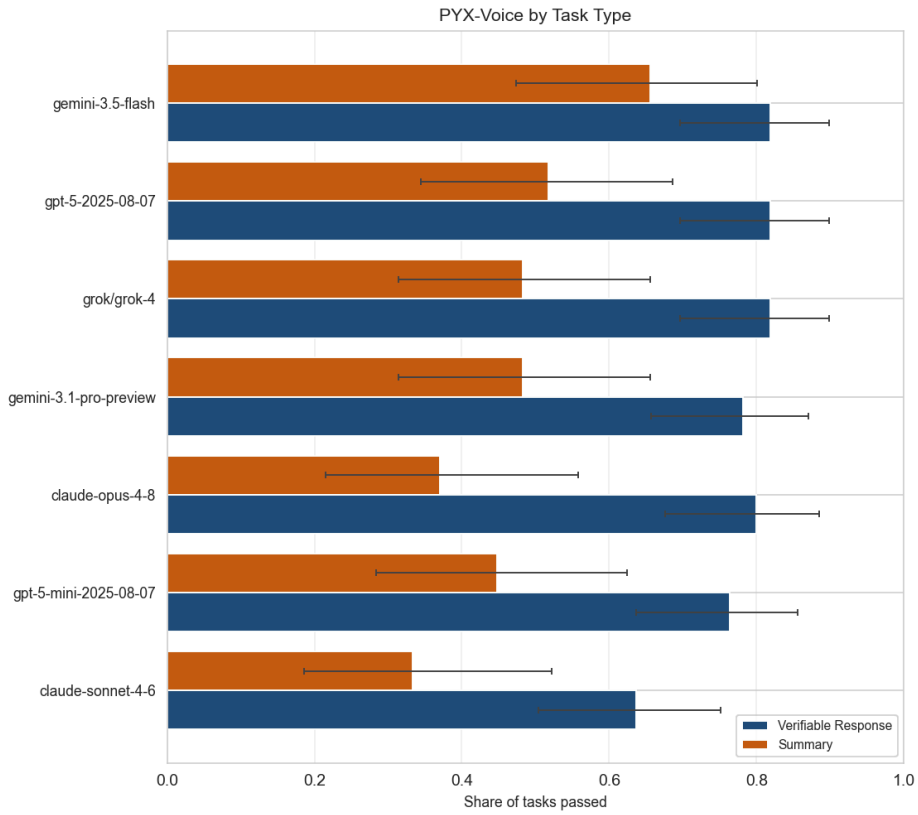
Figures

Figure 1. *Model performance across the entire benchmark. Values reflect the share of all 84 tasks passed.*



#	Model	Leaderboard Score	95% CI min	95% CI max
1	gemini-3.5-flash	76%	66%	84%
2	gpt-5-2025-08-07	71%	61%	80%
3	grok/grok-4	70%	60%	79%
4	gemini-3.1-pro-preview	68%	57%	77%
5	claude-opus-4-8	66%	55%	75%
6	gpt-5-mini-2025-08-07	65%	55%	75%
7	claude-sonnet-4-6	54%	43%	64%

Figure 2. *Model performance by task type. Values reflect the share of tasks passed per task type. Verifiable Response (n = 55) tasks are typically based on quantifiable metrics, with a binary correct or incorrect response. Summary (n = 29) tasks are typically based on open-ended employee feedback, involving summarizing data to arrive at thematic conclusions.*



#	Model	Verifiable Response Score	95% CI min	95% CI max
1	gemini-3.5-flash	82%	70%	90%
2	gpt-5-2025-08-07	82%	70%	90%
3	grok/grok-4	82%	70%	90%
4	claude-opus-4-8	80%	68%	88%
5	gemini-3.1-pro-preview	78%	66%	87%
6	gpt-5-mini-2025-08-07	76%	64%	86%
7	claude-sonnet-4-6	64%	50%	75%

#	Model	Summary Score	95% CI min	95% CI max
1	gemini-3.5-flash	66%	47%	80%
2	gpt-5-2025-08-07	52%	34%	69%
3	gemini-3.1-pro-preview	48%	31%	66%
4	grok/grok-4	48%	31%	66%
5	gpt-5-mini-2025-08-07	45%	28%	62%
6	claude-opus-4-8	37%	22%	56%
7	claude-sonnet-4-6	33%	19%	52%

Figure 3. *Model performance by LLM capability assessed. Values reflect proportion of tasks passed per capability, with higher values representing better performance. Each capability is assessed in a minimum of 5 tasks.*

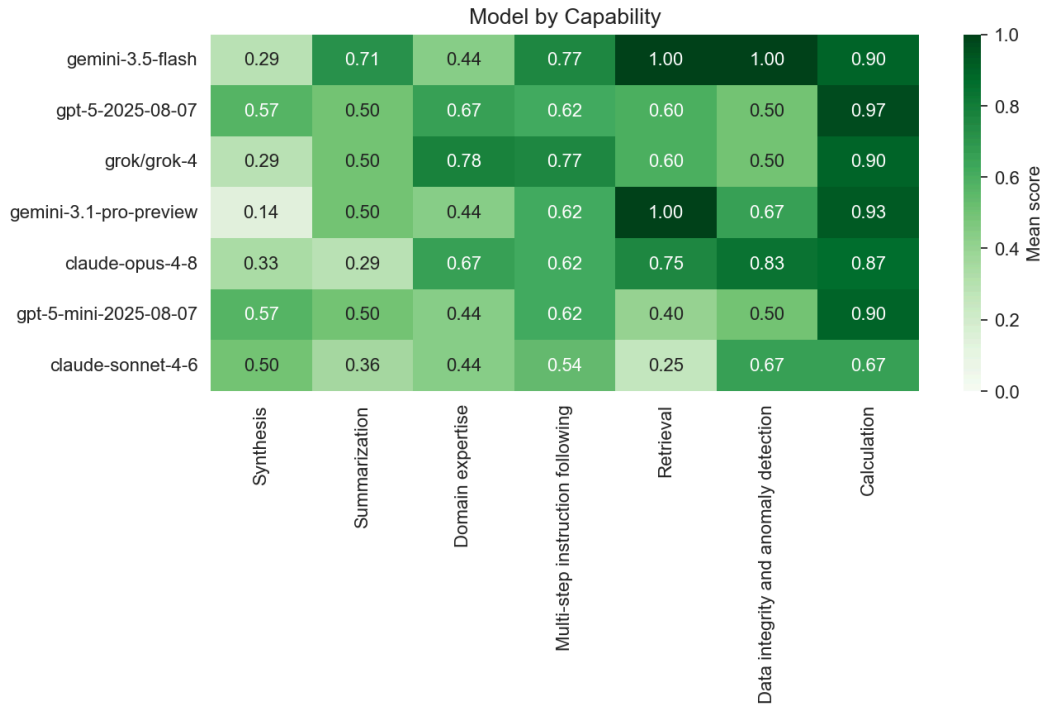


Figure 4. *Model performance by primary analysis type. Values reflect proportion of tasks passed per analysis type, with higher values representing better performance. *Indicates factors with fewer than 5 tasks. Results should be interpreted directionally.*

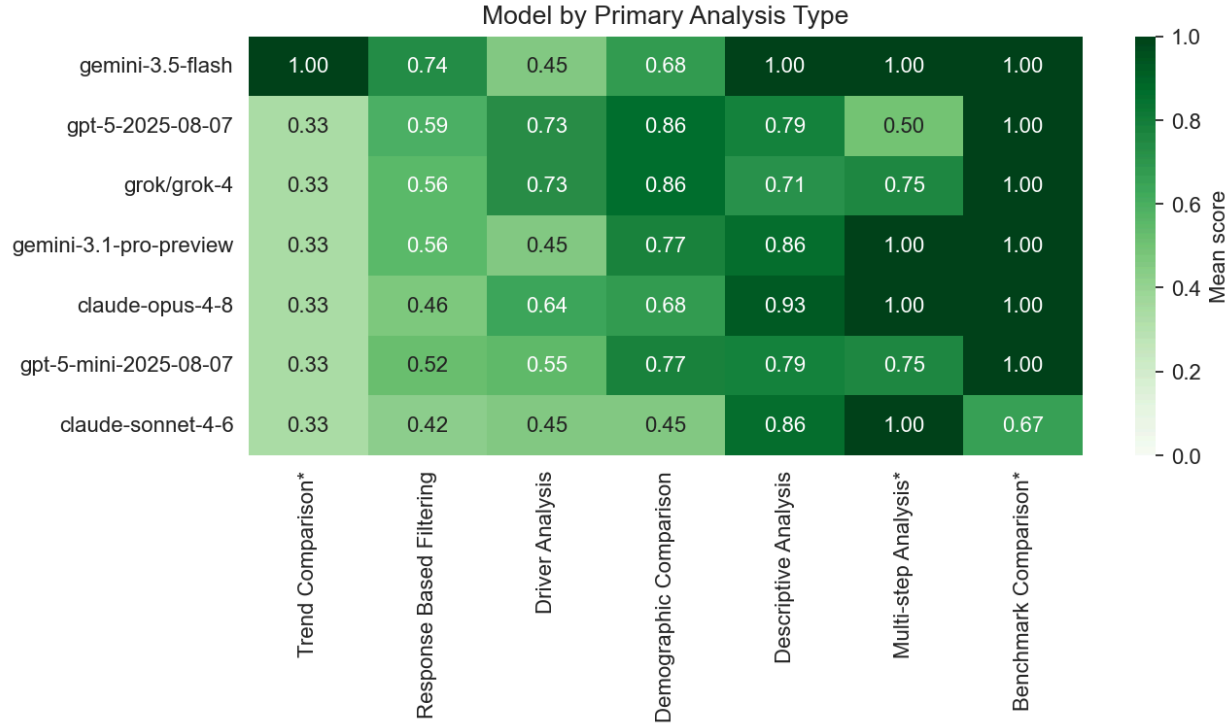


Figure 5a. *Percentage values reflect the mean scores across models for each employee experience factor in the Perceptyx People Insights Model. *Indicates factors with fewer than 5 tasks. Results should be interpreted directionally.*

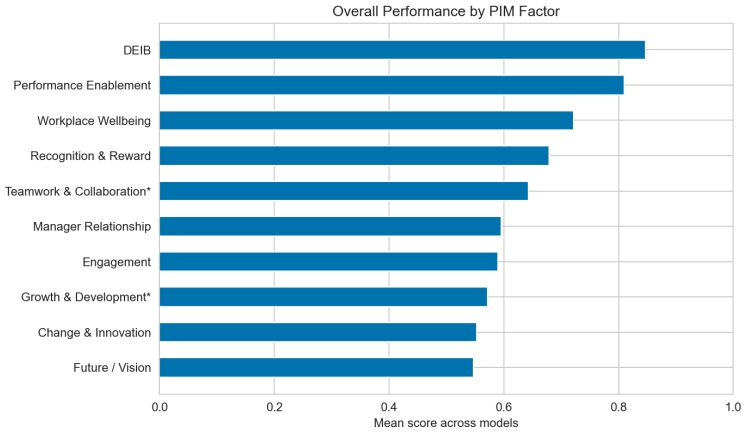


Figure 5b. *Values reflect proportion of tasks passed per employee experience factor, with higher values representing better performance. *Indicates factors with fewer than 5 tasks.*

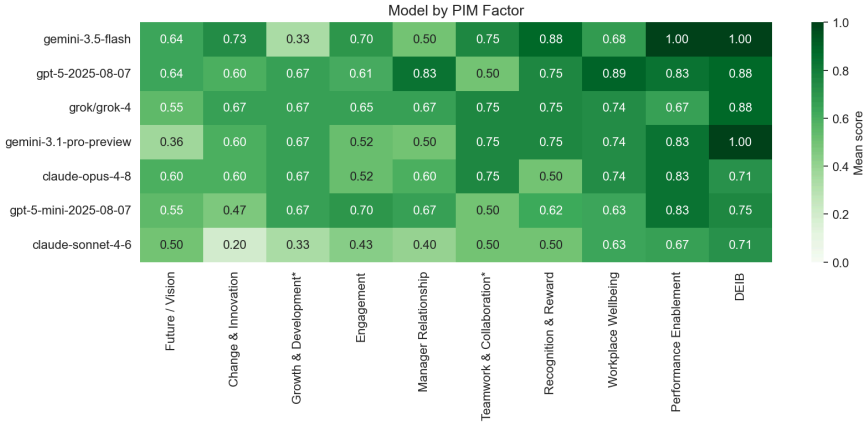
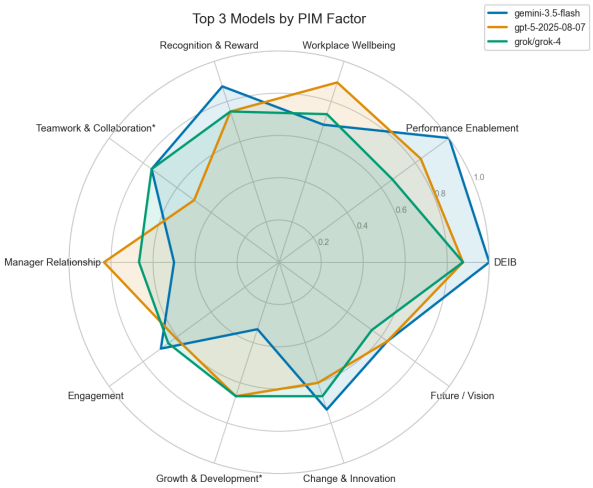


Figure 5c. *Top three performing models. Values reflect proportion of tasks passed per employee experience factor, with higher values representing better performance. *Indicates factors with fewer than 5 tasks.*



References

Burris, E. R. (2012). The risks and rewards of speaking up: Managerial responses to employee voice. Academy of Management Journal, 55, 851–875.

Burris, E., Thomas, B., Sodhi, K., & Klinghoffer, D. (2024, November). Turn employee feedback into action. Harvard Business Review. https://hbr.org/2024/11/turn-employee-feedback-into-action

Cramer, J. (2026, May 29). Company spent $500M on Claude in a single month: AI costs climb. Fast Company. https://www.fastcompany.com/91550884/claude-ai-costs-climb-company-spent-half-a-billion-dollars-in-a-single-month-report

Detert, J. R., Burris, E. R., Harrison, D. A., & Martin, S. R. (2013). Voice flows to and around leaders: Understanding when units are helped or hurt by employee voice. Administrative Science Quarterly, 58, 624–668.

Gans, J.S. (2023). Artificial intelligence adoption in a competitive market. Economica, 90(358), 690-705. doi: 10.1111/ecca.12458

Kossyva D, Theriou G, Aggelidis V, Sarigiannidis L. Outcomes of engagement: A systematic literature review and future research directions. Heliyon. 2023 Jun 22;9(6):e17565. doi: 10.1016/j.heliyon.2023.e17565. PMID: 37408885; PMCID: PMC10319181.

Lam, C. F., & Mayer, D. M. (2014). When do employees speak up for their customers? A model of voice in a customer service context. Personnel Psychology, 67, 637–666.

McClean, E. J., Burris, E. R., & Detert, J. R. (2013). When does voice lead to exit? It depends on leadership. Academy of Management Journal, 56, 525–548.

Olson, B. (2026, May 28). Corporate America is starting to ration AI as cost skyrockets. The Wall Street Journal. https://www.wsj.com/tech/ai/corporate-america-is-starting-to-ration-ai-as-cost-skyrockets-1eb99d7a

Perceptyx (2023, August 30). From insights to action: How AI is changing employee listening. https://blog.perceptyx.com/from-insights-to-action-how-ai-is-changing-employee-listening

Perceptyx (2026, April 15). Employee engagement: Why it matters and how to improve. https://blog.perceptyx.com/why-is-employee-engagement-important

Sajadieh, S., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Santarlasci, L., Pava, J., Maslej, N., Altman, R., Brynjolfsson, E., Brodley, C., Clark, J., Dignum, V., Kumar, V., Landay, J., Lyons, T., Manyika, J., Niebles, J.C., Shoham, Y., Tabassi, E., Wald, R., Walsh, T., Weld, D.. “The AI Index 2026 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2026.

Singla, A., Sukharevsky, A., Hall, B., Yee, L., & Chui, M. (2025, November 5). The state of AI: Global survey 2025. QuantumBlack AI by McKinsey. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/

Society for Human Resource Management (SHRM). (2026, March 31). The state of AI in HR 2026. https://www.shrm.org/topics-tools/research/state-of-ai-hr-2026

Zapata-Phelan, C. P., Colquitt, J. A., Scott, B. A., & Livingston, B. (2009). Procedural justice, interactional justice, and task performance: The mediating role of intrinsic motivation. Organizational Behavior and Human Decision Processes, 108, 93–105.

Appendix

Figure A1. *Critical-Failure Breakdown*

Model	Critical Failure	Count
gemini-3.5-flash	overclaim	2
gpt-5-2025-08-07	overclaim	2
grok/grok-4	overclaim	1
gemini-3.1-pro-preview	overclaim	2
claude-opus-4-8	overclaim	2
gpt-5-mini-2025-08-07	overclaim	2
claude-sonnet-4-6	overclaim	2
claude-sonnet-4-6	fabricated statistic	1

Figure A2. *Cost & Latency*

Model	Median duration (s)	Tokens in	Tokens out	Items
claude-opus-4-8	48.874	405,984	621,091	84
claude-sonnet-4-6	132.19	468,515	1,926,441	84
gemini-3.1-pro-preview	24.324	1,619,460	114,099	84
gemini-3.5-flash	50.109	4,852,368	237,560	84
grok/grok-4	24.216	999,887	111,112	84
gpt-5-2025-08-07	51.852	582,060	462,460	84
gpt-5-mini-2025-08-07	39.622	434,581	331,383	84