CALIBER: Calibrated Assessment via LLM for Interview Behavioral Evaluation and Rating

Katherine Kappaz **Raymond Alavo*** **Sofía Calderón** **Ahmed Abbasi**
Human-centered Analytics Lab, University of Notre Dame
salavo@nd.edu
https://caliber-assess.org

Abstract

We present **CALIBER**, an interactive dashboard for calibrated, LLM-assisted behavioral interview evaluation. The system is built around three core design principles. First, a structured **Annotate–Rubric–Rate (ARR)** workflow sequences each evaluation into span annotation, rubric consultation, and scoring—decomposing a holistic judgment into auditable steps. Second, LLM-generated STAR span highlights serve as *attention scaffolds* rather than scores: the model surfaces structural cues in the transcript while all final ratings remain the annotator's responsibility. Third, a 21-step guided training tour with Behaviorally Anchored Rating Scales (BARS) brings every evaluator to a shared baseline before live annotation begins. On the backend, a shrinking-pool randomizer with atomic DynamoDB counters guarantees uniform item coverage across crowdworkers, and a rich telemetry stream captures click, mouse-movement, and timing data alongside every annotation. A user study ($N = 24$) provides initial evidence consistent with the ARR design goals: both AI-friendly and AI-averse annotators complete tasks with less effort under AI assistance, and alignment with a held-out LLM reference improves for all groups. We release CALIBER[1] as an open-source toolkit together with a live demo[2] and code[3].

1 Introduction

Large language models are increasingly applied to high-stakes assessment tasks, including automated scoring of behavioral interviews—structured interviews in which candidates describe past experiences. Assessing behavioral interview responses is difficult for three reasons. **(a) Subjectivity:** raters weight STAR dimensions and competencies differently based on implicit schemata, producing divergent judgments even among trained evaluators (Highhouse, 2008). **(b) Lack of training:** many interviewers lack exposure to structured evaluation methods, leading to inconsistent rubric application (Barclay, 2001). **(c) Bias:** open-ended narrative responses invite evaluator bias independent of response quality (Langhan et al., 2023). These factors complicate the reliable ground-truth labels LLM-based scoring requires for calibration and evaluation.

Full automation also introduces circularity: without independent human labels, validating an LLM judge requires agreement with human references, not more model output (Zheng et al., 2023; Abbasi et al., 2024). It also conflicts with practitioner attitudes: although 88% of companies use AI in candidate screening (World Economic Forum, 2025), HR adoption remains uneven (Society for Human Resource Management, 2026), hindered by algorithm aversion—practitioners resist delegating decisions to algorithms even while acknowledging efficiency gains (Dargnies et al., 2024; Mirowska and Mesnet, 2022), preferring human intuition over algorithmic opacity (Highhouse, 2008; Mirowska and Mesnet, 2022). Overcoming this requires augmentation that does not increase workload or harm outcomes (Bankins et al., 2024; Prikshat et al., 2023; Noordeen, 2024). No existing tool supports this kind of human-in-the-loop augmentation. General-purpose annotation platforms like Prodigy (Montani and Honnibal, 2018), Doccano (Nakayama et al., 2018), LightTag (Perry, 2021), and INCEpTION (Klie et al., 2018) support span labeling or document classification, but none provide multi-dimensional rubric scoring, structured rater training, or LLM-assisted span scaffolding.

We present **CALIBER**, a web-based system for LLM-assisted behavioral interview evaluation built around an **Annotate–Rubric–Rate (ARR)** work-

*Corresponding author.
[1] https://caliber-assess.org
[2] https://youtu.be/fRpNjmASnpA
[3] https://github.com/nd-hal/CALIBER

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 466–475
October 24-29, 2026 ©2026 Association for Computational Linguistics

flow: evaluators review Situation–Task–Action–Result (STAR) evidence spans (Campion et al., 1997), consult structured criteria, and independently rate STAR quality, competency via Behaviorally Anchored Rating Scales (BARS), and STAR presence. CALIBER places AI assistance at evidence identification rather than final rating, keeping evaluators responsible for all final ratings.

Our main contributions are:

1. The **ARR framework**—Annotate, Rubric, Rate—a theoretically grounded interface design principle that sequences span annotation before scoring to promote consistent, evidence-based judgment.

2. A **multi-task annotation wizard** implementing ARR across four tasks: span extraction, structural scoring, BARS competency rating, and binary presence checklist.

3. **LLM-guided span scaffolding** via `claude-opus-4-8`, enabling a controlled study of whether attention guidance improves human rating quality.

The remainder of this paper describes related work (§2), the CALIBER interface (§3), system architectures (§4), and evaluation (§5).

2 Related Work

2.1 Behavioral Interview Assessment

Structured behavioral interviews ask candidates to narrate past experiences as evidence of competency, offering higher predictive validity than unstructured formats (Huffcutt and Arthur, 1994; Campion et al., 1997). Assessment quality nonetheless remains variable: assessors weight STAR dimensions differently based on implicit schemata (Highhouse, 2008), many lack formal exposure to structured evaluation methods (Barclay, 2001), and qualitative judgment invites biases independent of response quality (Langhan et al., 2023). These sources of variance motivate CALIBER's training-intensive, evidence-first design.

An key aspect of behavioral interview assessment is structured and rubric-based rating. BARS (Smith and Kendall, 1963) reduce subjective drift by tying scale points to concrete behavioral exemplars rather than abstract trait labels. STAR (Campion et al., 1997) operationalizes this at the response level, decomposing answers into four verifiable components. Together they underpin CALIBER's annotator training and multi-dimensional scoring design.

2.2 LLM-assisted Human Evaluation

MT-Bench, Chatbot Arena (Zheng et al., 2023), and Prometheus (Kim et al., 2024) established LLM-as-judge as a scalable evaluation paradigm, but trustworthy judges must be validated against high-quality human references (Alavo et al., 2026; Lalor et al., 2025) —exactly what CALIBER produces for the behavioral interview domain. How span-level LLM scaffolding affects rating consistency, particularly when scores are withheld to avoid anchoring, remains understudied.

With respect to behavioral interview NLP, Naim et al. (2018) examined multimodal cues in asynchronous video interviews. Benchmark construction for LLM-based interview scoring has lagged due to the absence of suitable annotation infrastructure—a gap CALIBER directly addresses.

2.3 General-purpose and Domain-specific Annotation Tools

The brat tool (Stenetorp et al., 2012) established web-based collaborative text annotation; Prodigy (Montani and Honnibal, 2018), Doccano (Nakayama et al., 2018), LightTag (Perry, 2021), and INCEpTION (Klie et al., 2018) extended it with active learning, task variety, and team management, while domain-specific tools extend it further to dialogue and clinical ontologies. All share a limitation for our use case: annotation as a single task per document, with no built-in multi-dimensional scoring, rater training flow, or LLM scaffolding.

3 The CALIBER Interface

CALIBER is a guided evaluation environment that connects transcript evidence, rubric guidance, and rating controls in a single, sequenced workflow. On arrival, annotators set an audio-narration preference and fill a background survey covering personality, HR experience, and demographics before encountering any interview material. The session is structured as a progressive calibration sequence: annotators first complete a **baseline evaluation** without any AI assistance—manually identifying STAR spans and rating the response—which serves as a calibration exercise that grounds their judgment in

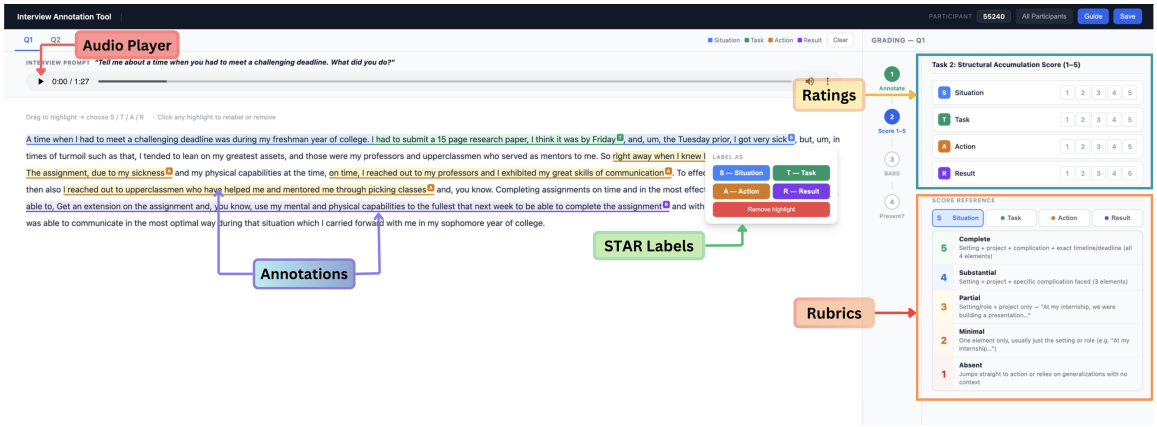


Figure 1: Screenshot overview of the CALIBER platform. Displaying the Annotate-Rubric-Rate (ARR) framework.

the rubric before AI augmentation is introduced. Once the baseline evaluation is complete, the system transitions to **AI-assisted mode**, where LLM-generated STAR span highlights are pre-populated and editable. In both stages all rating tasks are identical and all final scores are annotator-generated.

3.1 STAR Training and Guided Tutorial

CALIBER includes a built-in training sequence, run once per annotator before any live item, in two stages.

Conceptual tutorial. A six-step, audio-narrated module introduces the STAR framework (Campion et al., 1997) and the Behaviorally Anchored Rating Scale (Smith and Kendall, 1963), pairing each step's definition with a worked example.

Interactive guided tour. The training centerpiece is a **21-step interactive tour** on a fixed "Example" response, spotlighting each interface region with tooltip explanation and optional narration through a complete ARR cycle. The tour operationalizes Frame-of-Reference (FOR) training (Barclay, 2001): annotators encounter the same behavioral exemplars, developing a shared interpretive schema before working with real items (Figure 1).

3.2 Annotation–Rubric–Rating Interface

The core interface (Figure 1) is organized around the **ARR framework**: a task-sequencing principle enforcing evidence-gathering before evaluation. Rather than presenting a blank transcript and rating form simultaneously—which invites holistic, first-impression judgments (Highhouse, 2008)—ARR locks rating controls until each prior step is complete:

1. **Annotate.** The annotator drag-selects transcript spans for each STAR frame (Situation, Task, Action, Result); spans are color-coded and overlapping selections are permitted. This commits the annotator to specific textual evidence before any score is entered.

2. **Rubric.** A side panel displays the full rubric definition for each STAR frame and competency level, allowing on-demand reference without leaving the annotation view.

3. **Rate.** With evidence identified and rubric consulted, the annotator assigns (i) a 1–5 structural accumulation score per STAR frame, (ii) a single 1–5 BARS competency score, and (iii) a binary presence checklist (yes / no) for fine-grained criteria.

3.3 LLM-Assisted Span Annotation

In AI-assisted mode, the transcript is pre-populated with color-coded STAR span highlights generated offline by `claude-opus-4-8`. The model receives a prompt containing the full per-frame rubric and returns character-offset span boundaries for Situation, Task, Action, and Result (Appendix A). Highlights are injected as `<span class="hl hl-{s,t,a,r}">` markup and rendered with the same color scheme as manually created spans.

All LLM-generated spans are **fully editable**: annotators may extend, shrink, remove, or add spans before proceeding to rating. This preserves human agency over the evidence record while eliminating the blank-slate cost of span identification from scratch.

Crucially, *model scores are never shown to annotators*. Only span boundaries are surfaced; exposing numeric scores would risk anchoring ratings to the model's judgment rather than the annotator's

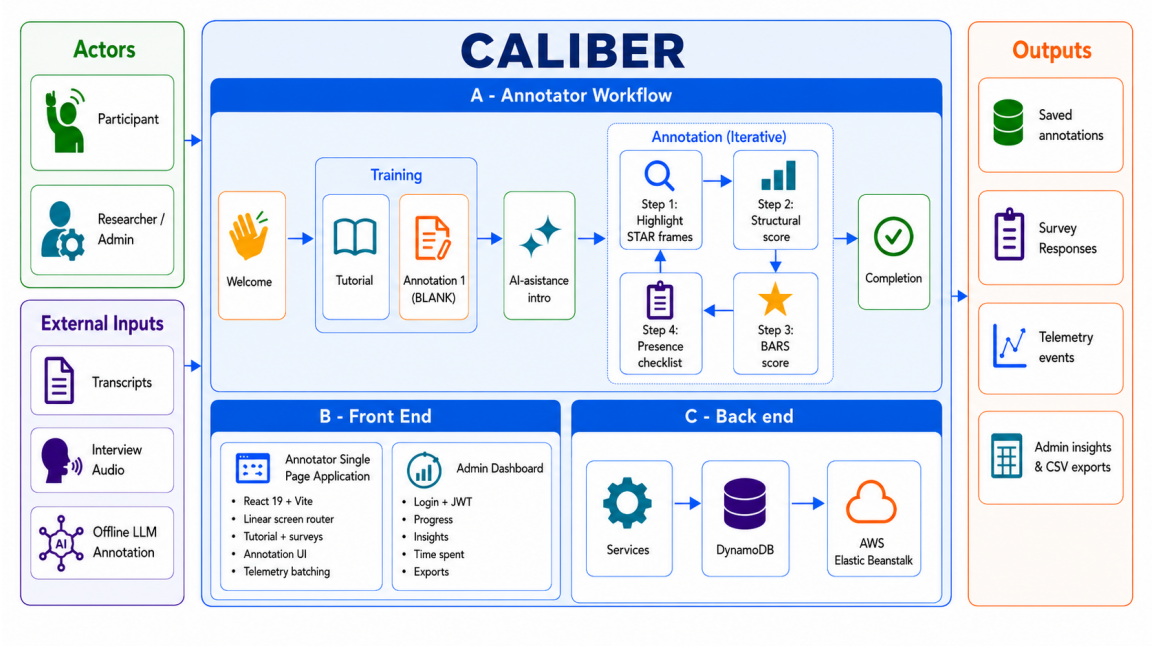


Figure 2: System design of CALIBER.

own reading of the transcript. BARS scores, structural scores, and binary presence labels are computed via a separate offline prompt (Appendix A) and held out as a reference for alignment analysis (Section 5.4).

4 System Overview

Figure 2 illustrates CALIBER's annotation pipeline. The system accepts three inputs: interview transcripts drawn from the participant item pool, audio recordings streamed from Amazon S3, and offline LLM-generated STAR annotations produced by `claude-opus-4-8` before the study begins. These inputs feed a four-step annotation pipeline guiding each annotator through the ARR workflow, persisting three outputs per session: annotation and score records, annotator profile data, and a continuous telemetry event stream. Table 1 summarizes the major pipeline components.

Component	Description
Inputs	Interview transcripts (participant item pool), audio recordings (S3), offline `claude-opus-4-8` STAR span annotations and scores.
Offline LLM	`claude-opus-4-8` generates STAR phrase spans, structural scores, BARS, and presence labels; stored read-only for hydration.
Annotation pipeline	Four-step ARR wizard: STAR span annotation, structural score (1–5), BARS competency rating (1–5), binary presence checklist.
Item assignment	Shrinking-pool randomizer with atomic DynamoDB conditional writes.
User actions	Onboarding → STAR training → 21-step guided tour → annotations → completion.
Outputs	Annotations CSV (spans + scores), annotators CSV (survey + milestones), telemetry CSV (clickstream events).

Table 1: Summary of the CALIBER annotation pipeline.

4.1 System Architecture and Implementation

CALIBER is implemented as a React 19 single-page application built with Vite, served by an Express/Node.js API deployed on AWS Elastic Beanstalk. Persistent state is distributed across multiple DynamoDB tables. Audio recordings are stored in S3 and delivered via presigned URLs.

Item ingestion and assignment. All the interview responses are saved into our database and processed offline by `claude-opus-4-8`. At session start, the **shrinking-pool randomizer** draws items for each annotator: it scans eligible items, applies Fisher-Yates shuffle, then claims items via atomic conditional calls. The first assigned item is delivered as part of the training (no LLM highlights); subsequent items are *AI-assisted* with pre-loaded span highlights injected as `<span>` markup.

Interaction telemetry. The `telemetry.js` module buffers UI events client-side and flushes them

in batches. Recorded events include click coordinates, mouse-movement traces at 1 Hz, and screen and phase transitions with millisecond timestamps. Telemetry can be disabled at deployment time for privacy-sensitive contexts.

5 Evaluation of CALIBER

CALIBER is evaluated in a behavioral interview assessment scenario comparing non-LLM-assisted and LLM-assisted annotation workflows within the same participants.

Interview and participant task. Each candidate answers two controlled behavioral interview questions (two minutes each): *"Tell me about a time when you had to meet a challenging deadline. What did you do?"* (before STAR training) and *"Tell me about a time when you had too many things to do and you were required to prioritize your tasks"* (after STAR training), standardizing difficulty while preserving the open-ended nature of behavioral responses. For each response, participants complete four tasks: (1) STAR span annotation, (2) per-frame structural rating (1–5), (3) BARS competency rating (1–5), and (4) binary STAR presence judgment.

Experimental flow. After background surveys, STAR training, and a 21-step guided tour, participants complete a **baseline** evaluation (manual span identification) followed by an **LLM-assisted** evaluation with editable `claude-opus-4-8`-generated span highlights pre-populated in the transcript. SUS is collected after each condition, alongside a post-task human–AI interaction battery covering automation desire and human agency (Shao et al., 2025) (8 items, 1–5 anchors), AI attitude, openness to AI suggestions, and reliance on AI (Noy and Zhang, 2023; Brynjolfsson et al., 2025; Dell'Acqua et al., 2023) (16 items, 7-point anchors), and GAAIS (Park et al., 2024) for AI-sentiment clustering.

LLM reference data. Before the study, each response is processed through `claude-opus-4-8` to produce reference STAR spans, structural scores, BARS scores, and presence judgments; only spans are surfaced to participants, while scores are withheld for alignment analysis. This isolates the effect of attention scaffolding from the anchoring that would result from revealing model scores.

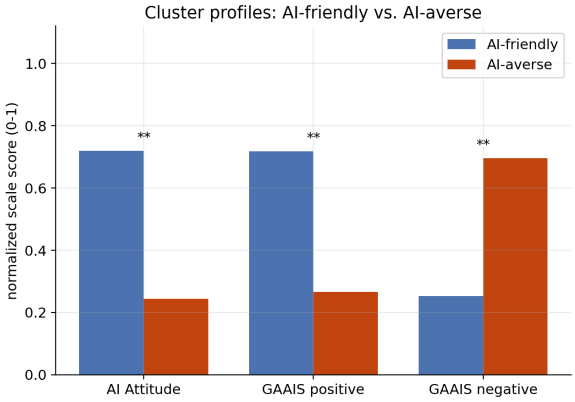


Figure 3: Participant profiles on AI-sentiment scales (used for clustering).

5.1 AI Sentiment Clustering

We cluster participants into two groups using k-means ($k = 2$) on three z-scored scales: *AI Attitude* (1–7), *GAAIS Positive* (1–5), and *GAAIS Negative* (1–5). The solution achieves a silhouette coefficient of 0.49, yielding 20 participants with strongly positive AI attitudes and 4 with negative ones. Following the convention in the human–AI interaction literature of *AI-appreciative* and *AI-averse* (Park et al., 2024), we adopt the shorthand **AI-friendly** and **AI-averse** throughout this paper for conciseness. Figure 3 shows the group profiles. The two clusters differ sharply on all three AI-sentiment scales (Mann-Whitney $p < 0.01$ for each), while being statistically indistinguishable on the two held-out SALT items: Automation Desire ($p = 0.25$) and Human Agency ($p = 0.94$). The partition reflects attitude toward AI specifically, not general task disposition.

5.2 Interaction Efficiency and Usability

Figure 4 reports time on task, click count, mouse-movement events, and SUS usability score by condition and group. **Both groups spend less effort when AI-assisted**: fewer clicks ($p < 0.01$ within each group), less time, and less mouse movement. Perceived usability (SUS) is unchanged between conditions and between groups. There are no significant between-group differences in any metric: AI-averse participants are not penalized relative to AI-friendly ones in either condition, consistent with CALIBER's design goal of avoiding workload costs for skeptical users. Because the baseline condition always precedes the AI-assisted condition, these efficiency gains are partly confounded with presentation order and item, and should not be read

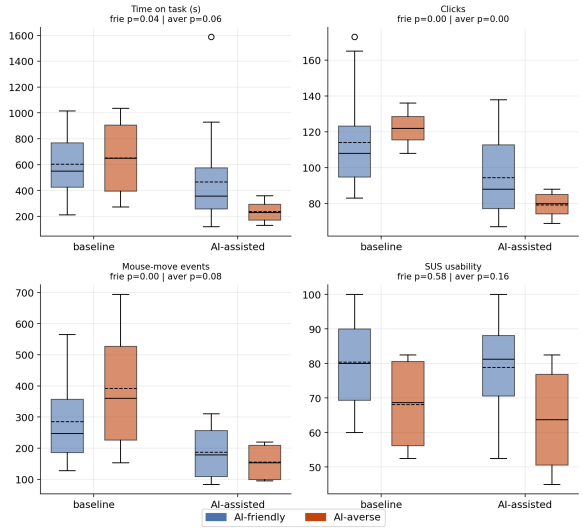


Figure 4: Effort and usability by group and condition. Both groups are faster, click less, and move the mouse less in the AI-assisted condition; SUS has no significant between-group differences.

as a clean causal estimate of scaffolding.

5.3 Annotation Interface Interaction

Mouse-cursor heatmaps and movement trajectories (Figure 5) show that participants engage with all regions of the ARR workflow regardless of condition. In both views, three activity zones emerge. ***Zone 1***, corresponding to the transcript panel (**Annotate**). ***Zone 2***, corresponding to the scoring panel (**Rate**) and the reference panel (**Rubric**), with a dense click and mouse concentration spanning across both panels. The presence of those two zones covering our three areas (Annotate - Rubric - Rate) across conditions is consistent with CALIBER's three-panel layout scaffolding the intended annotation sequence. We note that aggregate spatial density does not establish temporal ordering: heatmaps show *where* attention was allocated, not that participants visited the three stages in the prescribed sequence. Establishing sequence adherence would require analysis of ordered event transitions, which we leave to future work.

The conditions differ in *how* participants engage with the Annotate panel. In baseline, the transcript hotspot is larger and more diffuse, with long cursor traversals criss-crossing the text—indicating open-ended search for STAR-relevant segments. In the AI-assisted condition, the hotspot is tighter and more localised: LLM-generated span highlights concentrate attention on pre-identified segments, reducing exploratory scanning, and trajectories are correspondingly shorter and more directed. This shift is consistent with the efficiency gains in §5.2 (fewer clicks, less mouse movement) and suggests span scaffolding restructures the cognitive task from *search* to *verification*.



(a) Baseline. (b) CALIBER.

Figure 5: Mouse-cursor heatmaps and movement trajectories overlaid on the CALIBER interface: baseline (left) and AI-assisted (right).

5.4 Human–LLM Alignment

	AI-friendly (n=80)		**AI-averse** (n=16)	
Metric	Base. (n=40)	AI-asst. (n=40)	Base. (n=8)	AI-asst. (n=8)
$r \uparrow$	0.49	0.50	0.04	0.35
RMSE $\downarrow$	1.18	1.10	1.50	1.30
QWK $\uparrow$	0.48	0.50	0.04	0.33
$\kappa \uparrow$	0.54	0.70	−0.05	0.00

Table 2: Human–LLM alignment by group and condition. Per-condition n shown in column headers. $\uparrow$ higher is better; $\downarrow$ lower is better.

What alignment does and does not measure. The metrics below quantify *agreement with the LLM judge*, not annotation accuracy. Because the reference scores come from the same model (`claude-opus-4-8`) that generated the span scaffolding, increased alignment may reflect either improved judgment or convergence toward the model's framing; the present design cannot distinguish these. Human–human inter-rater reliability, the benchmark quality signal this work ultimately targets, is not estimable here, since each item receives a single rater per condition. We treat alignment as a proxy signal and return to this limitation in §6.

We measure alignment between participant STAR+BARS ratings and the LLM-as-a-judge using four complementary metrics. **Pearson** r captures linear correlation between human and LLM scores. **RMSE** (Root Mean Square Error) quantifies the average magnitude of rating disagreement on the original scale. **QWK** (Quadratic Weighted Kappa) measures agreement on an ordinal scale while penalising large disagreements more than small ones. **Cohen's** κ corrects for chance agreement between raters. Metrics are computed at the

question level ($N = 96$ total: 24 participants $\times$ 2 conditions $\times$ 2 questions); $n = 80$ for AI-friendly and $n = 16$ for AI-averse across both conditions ($n = 40$ and $n = 8$ per condition, respectively). Table 2 summarizes the results.

AI-friendly participants are already moderately aligned at baseline ($r = 0.49$, QWK$= 0.48$, $\kappa = 0.54$) and improve with AI assistance ($r = 0.50$, QWK$= 0.50$, $\kappa = 0.70$; RMSE $1.18 \rightarrow 1.10$).

AI-averse participants show a larger directional shift: unaided ratings are essentially uncorrelated with the judge (baseline $r = 0.04$, QWK$= 0.04$, $\kappa = -0.05$) and rise to moderate agreement under AI assistance ($r = 0.35$, QWK$= 0.33$; RMSE $1.50 \rightarrow 1.30$). This is suggestive that span scaffolding may matter most for the participants most resistant to AI augmentation, the population where consistency gains would be most valuable.

This result is exploratory. The AI-averse cluster contains $n = 4$ participants. All metrics in Table 2 are computed over question-level observations, which treats the four responses from each participant as independent and therefore overstates the effective sample size; participant-level statistics and confidence intervals are not estimable at this n. We report the pattern because it motivates a targeted follow-up study, not as a supported effect. It should not be read causally or used as a basis for design decisions without replication on a larger AI-averse sample.

6 Conclusion

We present CALIBER, an interactive dashboard for calibrated, LLM-assisted behavioral interview evaluation. The system's central contribution is the **ARR framework**—Annotate, Rubric, Rate—which sequences evidence identification before scoring to promote consistent, grounded judgment. Future work will add the capability for researchers to configure their own framework in order to personalize the behavioral interview structures, along with capabilities to personalize the scoring system and add new scoring metrics.

Limitations

Evaluation and Reliability Our primary quality signal, agreement with the LLM judge, is measured against the same model that produced the annotations provided in treatment setting. Disambiguating these requires an independent gold standard, such as expert-rater consensus. Similarly, producing trustworthy benchmark labels requires multiple independent judgments per item, but our within-subject design assigns each item to one rater per condition. Krippendorff's α, ICC, and related inter-annotator reliability metrics are consequently not estimable. Deploying CALIBER with at least two independent raters per item, and reporting reliability as a primary outcome, is the most important next step for the benchmark goal this work motivates.

Sample size, independence, and order confounds. The AI-averse subgroup comprises $n = 4$ participants. Reported metrics aggregate question-level observations, treating repeated measures within a participant as independent; this inflates apparent precision. The AI-averse result is labeled exploratory throughout and awaits replication. Furthermore, all participants complete the baseline condition before the AI-assisted condition, so assistance is confounded with presentation order, practice effects, and item identity. A counterbalanced or between-subjects design is needed to isolate the scaffolding effect.

No component ablation. The study contrasts assistance present versus absent. It does not isolate the individual contributions of ARR's components, evidence-based span annotation, the rubric panel, and rating sequencing, nor does it include a human+AI baseline without ARR structure.

Single model, single domain. All scaffolding and reference scores come from one model (`claude-opus-4-8`) applied to behavioral interview transcripts collected at a single institution. Generalization to other LLMs, interview formats, or rubric-based assessment domains requires separate validation.

7 Acknowledgments

This research was partially funded by an Industry Labs Seed Grant.

Ethics Statement

This work, including the user study, was reviewed and approved by the University of Notre Dame IRB (Protocol #25-10-9675). All Prolific participants provided informed consent prior to participation; the consent form explicitly describes the annotation tasks, and participants' right to withdraw at any time. No personally identifiable information beyond Prolific IDs is stored.

References

Ahmed Abbasi, Jeffrey Parsons, Gautam Pant, Olivia R Liu Sheng, and Suprateek Sarker. 2024. Pathways for design research on artificial intelligence. *Information Systems Research*, 35(2):441–459.

Raymond Alavo, Xinyuan Zhang, Gemza Ademaj, Junhui Cai, Hyeokhyen Kwon, Robert Cotes, Gari D Clifford, and Ahmed Abbasi. 2026. Dash-m5h: An interactive dashboard for multi-modal, multi-model mental health assessment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 138–147.

Sarah Bankins, Anna Carmella Ocampo, Mauricio Marrone, Simon Lloyd D. Restubog, and Sang Eun Woo. 2024. A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*, 45(2):159–182.

James M. Barclay. 2001. Improving selection interviews with structure: Organisations' use of 'behavioural' interviews. *Personnel Review*, 30(1):81–101.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2025. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942.

Michael A. Campion, David K. Palmer, and James E. Campion. 1997. A review of structure in the selection interview. *Personnel Psychology*, 50(3):655–702.

Marie-Pierre Dargnies, Rustamdjan Hakimov, and Dorothea Kübler. 2024. Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Scott Highhouse. 2008. Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1(3):333–342.

Allen I. Huffcutt and Winfred Arthur. 1994. Hunter and hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, 79(2):184–190.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Han, Chaehyeong Kwon, Sejune Park, and 1 others. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. *Preprint*, arXiv:2310.08491.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

John P Lalor, Ruiyang Qin, David Dobolyi, and Ahmed Abbasi. 2025. Textagon: Boosting language models with theory-guided parallel representations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–92.

Melissa L. Langhan, Diane E. J. Stafford, Angela L. Myers, Bruce E. Herman, Megan L. Curran, Angela S. Czaja, and Richard Mink. 2023. Clinical competency committee perceptions of entrustable professional activities and their value in assessing fellows: A qualitative study of pediatric subspecialty program directors. *Medical Teacher*, 45(6):650–657.

Agata Mirowska and Laura Mesnet. 2022. Preferring the devil you know: Potential applicant reactions to artificial intelligence evaluation of interviews. *Human Resource Management Journal*, 32(2):364–383.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A modern annotation tool for machine learning.

Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2018. Automated analysis and prediction of job interview performance. In *IEEE Transactions on Affective Computing*, volume 9, pages 191–204.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human.

Abdul Rahman Noordeen. 2024. Are employees avoiding or approaching AI? In *AMCIS 2024 TREOs*. Paper 96.

Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Soojin Park, Peter Kvam, and Claire Youngjin Kim. 2024. Attitudes toward artificial intelligence: Development and validation of a multidimensional scale. *International Journal of Human–Computer Studies*, 182:103168.

Tal Perry. 2021. LightTag: Text annotation platform. In *Proceedings of the 2021 Conference of the North American Chapter of the ACL: Demonstrations*.

Verma Prikshat, Mohammad Islam, Parth Patel, Ashish Malik, Pawan Budhwar, and Suraksha Gupta. 2023. AI-augmented HRM: Literature review and a proposed multilevel framework for future research. *Technological Forecasting and Social Change*, 193:122645.

Ruomeng Shao and 1 others. 2025. Future of work with AI: Automation desire and human agency in professional tasks.

Patricia Cain Smith and Lorne M. Kendall. 1963. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47(2):149–155.

Society for Human Resource Management. 2026. The state of AI in HR 2026. SHRM Thought Leadership.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at EACL 2012*, pages 102–107.

World Economic Forum. 2025. Hiring with AI doesn't have to be so inhumane. here's how. World Economic Forum Publications.

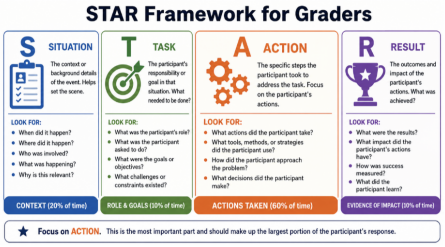
Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A LLM Annotation Prompt: STAR Text Extraction

CALIBER includes an LLM-based annotation pipeline designed to pre-populate STAR span highlights for the multi-task annotation wizard described in Section 3, with the goal of scaffolding participant attention without revealing model-generated scores. We use `claude-opus-4-8`, prompted to act as an experienced human resources professional evaluating behavioral interview transcripts. The model is instructed to extract, for each transcript, the exact quoted phrases corresponding to Situation, Task, Action, and Result, following the same definitions later shown to participants during STAR training (Section 3). Extracted spans may overlap within a single sentence (e.g., a candidate may state their Task and first Action in the same breath) and may range from a few words to a full paragraph in length. Figure 6 shows a worked transcript with its extracted Situation, Task, Action, and Result phrases. The full extraction prompt is shown below.

Prompt for STAR Text Extraction

Role: You are an experienced human resources professional. You will evaluate brief behavioral interview responses.
Task: Your goal is to extract the STAR elements from the candidate's story transcript. Read through the text and extract the exact words/phrases that correspond to the Situation, Task, Action, and Result.



Keep in mind:

- **Overlap is normal:** A single sentence might contain multiple labels (e.g., a candidate states their Task and their first Action in the same breath).
- **Length varies:** A single label might stretch across a whole paragraph, or it might be just five words long.

For "Situation," select text describing the background, setting or context of the story. For "Task," select text describing the specific problem, goal or deadline the candidate had to meet. For "Action," select text describing the specific observable steps the candidate took personally to solve the problem. For "Result," select text describing the outcome, impact or learning that occurred because of the candidate's actions.

Example transcript: "During an internship I did this past summer, we had a project where the deadline was in about a couple weeks and we had to coordinate between teams and not just the Chicago office, but also the Dallas and New York office. The project was essentially putting together a consulting presentation to help a nonprofit meet new goals and kind of improve their revenue for the upcoming year. And what I did to meet the deadline was essentially make sure to touch base with the team every day, whether over a meeting or just, you know, through our teams chat and made sure that every single person knew what they were, what they were gonna work on that day. I would also try to check in throughout each day and make sure that we were making progress and sort of on schedule to meet our upcoming deadline. And this

Um, last week in class I had to write a paper for my writing and rhetoric class and I only started one day before it was due. So in study hall, I sat down with my tutor and we made a plan together on how to tackle the paper and then I wrote like a rough draft and I met with one of the people at the writing center outside and we went, went over it and finalized it and that's how I got my paper done.

Figure 6: Example STAR text extraction. Situation, Task, Action, and Result phrases are highlighted directly in the transcript, including a case of overlapping labels within a single sentence.

was also kind of helpful to figure out where people needed help, where people might have kind of finished a little bit early and can potentially help someone else out. And yeah, ultimately that led to us being able to meet our deadline and with some extra time we had to, we were able to kind of, you know, go above and beyond a little bit as we seek feedback before we were even, you know, submitting and presenting to project manager."
Example extracted phrases:

- **Situation:** "During an internship I did this past summer"; "not just the Chicago office, but also the Dallas and New York office"
- **Task:** "we had a project where the deadline was in a couple of weeks"; "The project was essentially putting together a consulting presentation to help a nonprofit meet new goals"
- **Action:** "what I did to meet the deadline was essentially make sure to touch base with the team every day"; "made sure that every single person knew what they were, what they were gonna work on that day"; "I would also try to check in throughout each day and make sure that we were making progress"
- **Result:** "ultimately, that led to us being able to meet our deadline"; "we were able to kind of, you know, go above and beyond a little bit"

Output format: For each transcript in the spreadsheet, you should add four columns: one each for the S, T, A, and R related text phrases (e.g., `a1_S_phrases`). Place each phrase in quotes. If a given column has multiple phrases, separate them with a semi-colon.