Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 1046–1056
https://doi.org/10.54941/ahfe1006917

Can LLMs Assist in Job Interview Preparation? Assessing the Quality and Effectiveness of LLM-Generated Feedback

Ghritachi Mahajani, Amir Behzadan, and Theodora Chaspari

University of Colorado Boulder, CO 80309, USA

ABSTRACT

Large language models (LLMs) have demonstrated strong reasoning capabilities, making them ideal for generating formative feedback in learning contexts. This paper evaluates the ability of LLMs to provide formative feedback on interviewees' responses in a job interview task. Specifically, the degree of explanation in an interviewee's response, a key communication skill, is used as the key assessment criterion. Combinations of LLM models (i.e., GPT-3.5-Turbo, Gemini-1.5-Pro) with various chain-of-thought (CoT) prompting strategies, including task definition, domain knowledge, and contrastive prompting, are examined across multiple self-reported metrics of feedback quality effectiveness. Data is collected from 663 participants on Amazon MTurk using a between-subjects design. Results indicate that users perceived LLMs as having a moderate ability to provide formative feedback in job interviews, though the feedback was at times viewed as irrelevant or potentially offensive. It is also found that the choice of LLM model and prompting strategy significantly influences perceived feedback quality. While stronger task performance occasionally aligned with higher user ratings, the relationship between performance and perception is not strictly linear. Findings are discussed in terms of design implications for enhancing the quality and effectiveness of LLM-generated feedback in interview training.

Keywords: Large language models, Formative feedback, Job interview, Human-AI Interaction

INTRODUCTION

Formative feedback, defined as actionable information for learners (Shute, 2008), is crucial in personalized training across various domains (e.g., education, sports, healthcare), and can promote reflection and enhance performance. Research has explored formative feedback systems for teacher training (De Angelis and Miranda, 2023) and student learning (Xu et al., 2023), as well as for patients and medical experts (Polonsky and Fisher, 2015; Naik et al., 2018). In oral communication, feedback has typically been given through peer and self-reviews (Smith et al., 2020). In interview training, research has largely focused on demonstrating favourable social skills and personality traits via modifying non-verbal signals (e.g., smiles, head nods, body language), as well as social cues and emotional expressions (Anderson et al., 2013; Gebhard et al., 2018; Hoque et al., 2013). However, micro-level,

turn-by-turn feedback on linguistic aspects is underexplored, even though it plays a crucial role in communicating reasoning and confidence during interviews (Kalin and Rayko, 2013; Naim et al., 2015), by pinpointing which parts of an interview response are effective and which ones may require revision.

Large language models (LLMs) offer new capabilities in natural language understanding with minimal supervision, making them promising for spoken language understanding (SLU), especially where task-specific labeled data is limited (Aggarwal et al., 2025; He and Garner, 2023; Li et al., 2023). Beyond content detection, LLMs can reason through problems, which could enhance formative feedback. However, their performance in high-stakes, real-world scenarios requiring semantic inference and domain expertise remains unclear. LLM reasoning has primarily evaluated the correctness of the final answer rather than the reasoning process itself (Dougrez-Lewis et al., 2024). Important evaluation dimensions, e.g., perceived informativeness, understandability, and completeness of the generated responses by the LLM have largely remained overlooked (Carton et al., 2020).

This study evaluates the ability of LLMs to provide formative feedback on interviewee responses during job interviews by classifying the response's degree of explanation. The degree of explanation is selected as a key assessment criterion reflectnig communication ability and self-awareness as it captures how thoroughly and clearly an interviewee justifies or elaborates on their response (Levashina et al., 2014). We examine different models and prompting techniques, and investigate three research questions: **(RQ1):** To what extent can LLMs provide formative feedback for interview training? **(RQ2):** How do different LLM models and prompting techniques affect the perceived quality of the feedback? **(RQ3):** Is there a relationship between user perceptions of LLM feedback and the model's performance on the focal task? To address these questions, we conduct a user study on Amazon MTurk with a between-subjects design across six experimental conditions that correspond to combinations of different LLM types and prompting strategies. Participants evaluate feedback based on criteria such as understandability, informativeness, and agreement, and share their overall impressions of the system's trustworthiness and effectiveness.

Our contributions are: (1) Unlike prior work in interview training that evaluated user performance holistically, we emphasize turn-by-turn feedback based on individual responses, allowing for more actionable insights; (2) In contrast to the majority of studies on LLM reasoning that have been conducted on mathematical, commonsense, and logical tasks, we evaluate LLM reasoning on a specialized human-centered, real-world task requiring semantic and contextual understanding; and (3) We compare user perceptions of the quality and effectiveness of the LLM feedback across different LLM architectures and prompting techniques.

RELATED WORK

Rationalization in natural language processing (NLP) refers to the process of making language models more interpretable by generating justifications for their outputs, often in the form of natural language explanations

(Gurrapu et al., 2023). Self-rationalization leverages an LLM's own reasoning capabilities to explain its decisions, and was first demonstrated through CoT prompting that enables LLM reasoning through intermediate reasoning steps (Wei et al., 2022). CoT significantly improves task accuracy on various deductive reasoning tasks, even in zero-shot settings (Kojima et al., 2022). However, other types of reasoning, such as abductive, inductive, and causal reasoning, remain more challenging for LLMs and have been less extensively studied (Dougrez-Lewis et al., 2024). Lyu et al. (2024) examined the faithfulness of 110 model-generated explanations in NLP and highlighted concerns about the risks posed by unfaithful rationales, especially when they appear plausible (i.e., logically coherent and well-structured). Yet, other work indicates that even when an LLM generates a convincing explanation for an incorrect output or label, such explanations may still be valuable to users by sparking new ideas or prompting reflection (Okoso et al., 2025).

Prior work on evaluating LLM reasoning has utilized both automated methods to assess the utility and correctness of generated rationales, and user studies that gather human perceptions of these properties. Joshi et al. (2023) investigated the utility of machine-generated rationales for question answering and found that commonly used evaluation metrics, such as LLM task performance or the similarity between generated and gold-standard rationales, did not reliably predict human utility. Instead, features like conciseness and novelty are more indicative, though difficult to estimate without human input. Carton et al. (2020) proposed the use of fidelity curves to better characterize rationale quality in terms of sufficiency (i.e., ability to fully explain the output) and comprehensiveness (i.e., the extent to which a rationale is needed for a prediction). The study concluded that the concept of one-size-fits-all fidelity benchmarks is problematic, underscoring that human rationales should not be treated as gold standards and that careful procedures are needed to collect, understand, and interpret the properties of rationales and their evaluation metrics. Lubos et al. (2024) conducted an online study with 93 users to demonstrate the potential of LLMs to generate high-quality, personalized explanations that support users across different types of recommendation approaches. Results showed positive user perceptions of LLM-generated explanations that helped users assess the relevance and usefulness of the recommended items. Chen et al. (2024) evaluated the quality of LLM-generated explanations using human and automated evaluations.

PROMPT DESIGN AND GENERATION OF REASONING

Study data is obtained from audio recordings and transcripts of 38 mock interviews in the VetTrain dataset (Project VetTrain, n.d.). Collectively, the dataset contains 286 question-response pairs, in which the interviewer asked a question and the interviewee provided a response, possibly with some follow-ups. Three independent annotators labeled the degree of explanation of each response as one of four possible categories: 'under-explained' ($n = 23$), 'succinct' ($n = 107$), 'comprehensive' ($n = 122$), and 'over-explained' ($n = 24$) (Verrap et al., 2022). The disagreement among the annotators was resolved through multiple rounds of adjudication, resulting

in Krippendorff's $\alpha = 0.677$. For each response, the final label was obtained through majority voting.

We conducted text classification tasks in which the LLM was asked to provide a decision and its reasoning about the degree of explanation of a question-response pair. Taking into account that the responses belonging to the 'under-explained' and 'succinct' classes are significantly shorter in word length compared to ones from the 'comprehensive' and 'over-explained' classes ($t(216.52) = -14.07$, $p < 0.01$), we framed two binary classification tasks: (1) Short task, that included 130 question-response pairs from 'under-explained' and 'succinct' responses; and (2) Long task, that included 156 question-response pairs from 'comprehensive' and 'over-explained' responses. All prompts included an introductory sentence setting up the context for the LLM: "*You are a text classifier. Your task is to classify the following interview response into one of two categories*:".

Following that, we elicited the LLM decision of the degree of explanation and reasoning via three types of prompts: (1) *Definitions*: including the definition of the classes of the corresponding task and the prompt "*Give your response in the form label:<class>, reasoning: <reasoning>*"; (2) *Definitions + Domain knowledge (DK)*: including the definition of classes, domain knowledge regarding the psycholinguistic characterization of the degree of explanation in the responses, and the same final prompt as the previous one. As part of the DK, three descriptors (i.e., % words related to politics, tentativeness, politeness) were included in the prompt for the Short task, and 12 descriptors (i.e., word count, prepositions, % words related to numbers, achievement, causation, negations, tentativeness, social processes, auditory processes, work, past) were included for the Long task. These descriptors were extracted via the Linguistic Inquiry and Word Count (LIWC) toolbox (Pennebaker et al., 2015). The descriptors were converted into simple self-explanatory text and were added to the prompt; and (3) *Definitions + Why/Why Not CoT*: including the definition of the classes of each task and a CoT prompt that encourages the LLM to provide a 'why-choose' and 'why-not-choose' reasoning to explain why a specific degree of explanation was chosen and the other was dismissed (Chen et al., 2024). We experimented with both GPT-3.5-Turbo and Gemini-1.5-Pro models, considering all the above three types of prompts, resulting in six combinations corresponding to the conditions of the user study (Table 1).

Table 1: LLM configurations and performance of the six experimental conditions.

Conditions	Models	Description	F1-Score
C1	Gemini	Definitions	0.37
C2	GPT	Definitions	0.29
C3	Gemini	Definitions + DK	0.37
C4	GPT	Definitions + DK	0.49
C5	Gemini	Definitions + Why/Why Not CoT	0.55
C6	GPT	Definitions + Why/Why Not CoT	0.23

METHODS

Participant Recruitment and Study Protocol

Participants were recruited from MTurk, with inclusion criteria of being a U.S. resident, and having English proficiency, 98%+ HIT approval rating, and 1,000+ completed HITs. The study was distributed as a HIT on MTurk, directing participants to a Qualtrics survey. After accepting the HIT and providing informed consent, participants were randomly assigned to 1 of six conditions (Table 1), evaluated AI-generated feedback on 4–5 question-response pairs (either from the Short or Long task), and completed a series of post-case evaluation metrics for each pair followed by an overall perception of the system in a post-study survey. Upon completion, they received a unique code for compensation of $1 per HIT.

In total, we obtained 2,398 complete responses from Qualtrics. Checking for valid survey codes, attention check responses, and minimum timing constraints (finishing in < 2 min. was considered infeasible) disqualified 617, leaving 1,778 responses. Cleaning data based on duration of completion removed some noise from the values, reducing the size to 1,030 entries. On average, 3.16 participants rated each question-response pair. For each pair, ratings were averaged out, resulting in a total of 1,030 samples: 169 for C1, 182 for C2, 182 for C3, 170 for C4, 157 for C5, and 170 for C6. These samples corresponded to 663 participants (292 female, 371 male), most of whome were aged 25–34 ($n = 405$), followed by 35–44 ($n = 135$), 18–24 ($n = 41$), and smaller numbers in the 45-54, 55-64, and 65+ age groups. The majority held a bachelor's degree ($n = 527$), with several others reporting a graduate/professional degree ($n = 107$), and smaller numbers having associate/technical degree, high school diploma/general educational development (GED), some college but no degree, or some high school or less.

Measures

We used two types of questionnaires. The first was administered after each question-response pair (i.e., post-case), therefore it was completed 4–5 times per participant. It had 10 items related to common measures obtained in prior work associated with feedback evaluation from LLMs (Chen et al., 2024). Particularly, we measured on a 5-point Likert scale, the perceived *understandability* and *fluency* of the model reflecting the clarity of the LLM reasoning; *informativeness*, *relevance*, and *irrelevance* reflecting the degree of presence of valuable information in the reasoning; *transparency* for the extent to which users believe they understand the decision-making process of the LLM; *persuasiveness* and *effectiveness* for the extent to which participants would consider using the points in the LLM reasoning as feedback; *offensiveness* framed as the presence of discriminatory or offensive content in the reasoning; and *agreement* with the LLM serving as a proxy of the perceived quality of the feedback. The second questionnaire was administered after the participants had reviewed all cases (i.e., post-study), in order to rate, on a 5-point Likert scale, the extent to which they trust the system and believe that the system is effective at giving feedback.

METHODS AND RESULTS

We conducted a one-way ANOVA to assess significant differences across the six conditions regarding the considered perceptual metrics, followed by post hoc t-tests. The six conditions depicted significant differences in terms of informativeness ($F(5,1638)=7.15$, $p<0.01$), transparency ($F(5,1638)=3.85$, $p<0.01$), harm ($F(5,1638)=2.75$, $p<0.01$), and agreement ($F(5,1638)=4.82$, $p<0.01$). Figure 1 presents the t-test results. The C3 model demonstrated the lowest scores in both informativeness and agreement. The C1 model was perceived as the most transparent configuration, scoring significantly higher than C2 model ($t(349) = 2.00$, $p = 0.04$) and C3 model ($t(349) = 3.38$, $p < 0.001$). The C5 model was viewed as providing the least offensive feedback, significantly lower than C2 model ($t(337) = 3.07$, $p < 0.01$), C3 model ($t(337) = 3.01$, $p < 0.01$), C6 model ($t(325) = 2.60$, $p < 0.01$), C1 model ($t(324) = 2.35$, $p = 0.01$), and C4 model ($t(325) = 2.30$, $p = 0.02$). The C3 model was consistently the worst-performing condition, indicating that statistical domain knowledge was not used by Gemini effectively. Post-study trust did not show statistically significant differences across the conditions. In contrast, perceived effectiveness varied significantly across the conditions. The C5 model was rated as the most effective overall, with significantly higher effectiveness scores compared to C2 model ($t(1024) = 3.11$, $p < 0.01$), C4 model ($t(1024) = 2.45$, $p = 0.01$), and C3 model ($t(1024) = 2.14$, $p = 0.03$).

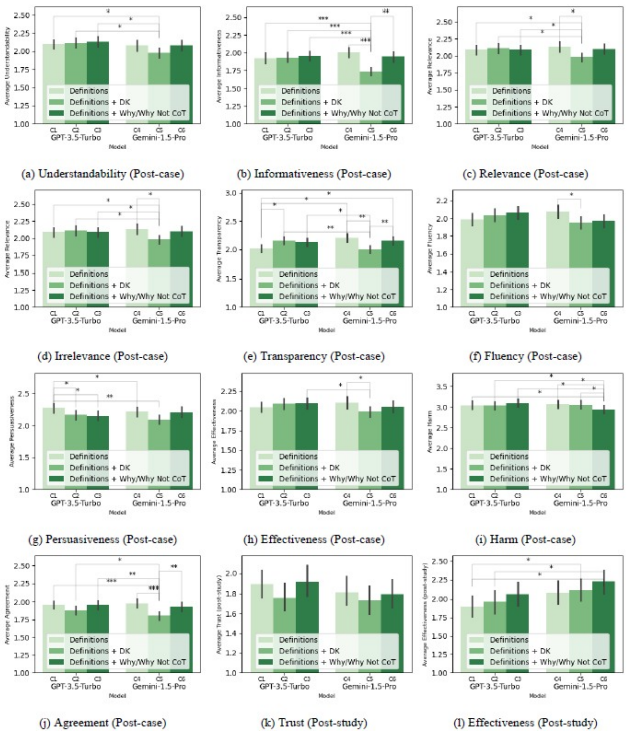

Figure 1: Bar plots of participants' perceived scores for the different configurations of models. The results from the post-hoc analysis of pairs of models via t-tests are also annotated (***: p<0.001, **: p<0.01, *: p<0.05).

To examine the relationship between LLM characteristics and user perceptions of feedback quality, we conducted a series of linear regression (LR) analyses. Given the limited sample size, we estimated four separate models (LR models 1-4) to reduce overfitting (Table 2). All models included two key independent variables: the type of LLM model (i.e., 0: Gemini, 1: GPT) and its task performance measured by F1-score (Table 1). To assess the influence of different prompting strategies, LR models 2 through 4 each included one of the three prompting techniques (i.e., "Definitions", "Definitions + DK", "Definitions + Why/Why Not CoT") as an additional independent variable. The presence of prompting strategy was coded as "1" and its absence as "0". Results indicated an overall participant preference for the GPT model over Gemini, particularly in terms of understandability and informativeness across all four LR models. The LLM F1-score did not show a significant effect for most perceptual metrics, with the exception of harm, where participants perceived models with lower F1-scores as more harmful. Prompting with "Definitions + CoT" was associated with increased informativeness, persuasiveness, and agreement. In contrast, prompting with "Definitions + DK+ CoT" had a negative impact, resulting in significantly lower perceived informativeness, transparency, persuasiveness, and agreement, along with significantly higher perceived offensiveness. Finally, the "Definitions + Why/Why Not CoT" prompt yielded a significant reduction in perceived offensiveness.

DISCUSSION

In answering (**RQ1**) on the extent to which LLMs can provide formative feedback for interview training, we found that different model combinations received moderate ratings overall. Understandability, transparency, fluency, informativeness, and effectiveness generally fell in the 1.72-2.12 range on a 5-point Likert scale, suggesting that users found LLM-generated feedback somewhat useful and coherent. In contrast, irrelevance and offensiveness received higher ratings, between 2.85 and 3.09, indicating that feedback was often perceived as off-topic or potentially offensive. Trustworthiness was rated in the 1.74-1.91 range, reflecting user skepticism of the models. However, effectiveness scores ranging from 1.90 to 2.23 suggest that user still saw some actionable value in the feedback. Overall, these findings indicate that while LLM-generated feedback holds promise, it might not yet meet the standards necessary for reliable interview training.

In answering (**RQ2**), our findings indicate that the choice of LLM and the prompting strategy significantly influences users' perception of feedback quality. The GPT model received more favorable ratings than the Gemini model, consistent with previous research indicating that GPT excels in generating relevant, complex, structured, and creative content (Lang et al., 2024). The "Definitions + DK" prompting strategy underperformed across several perceptual metrics, particularly when paired with Gemini. One potential reason for this is that participants may not have been familiar with the psycholinguistic domain knowledge embedded in the prompt, possibly

Table 2: Linear regression (LR) results predicting perceptual metrics on LLM feedback, including model fit statistics and coefficients (standard deviation in parentheses).

LR Model	Metric	LR Fit Statistics	LLM Model (0: Gemini, 1: Gpt)	LLM F1-Score	Definitions + CoT (0: No, 1: Yes)	Definitions + DK + CoT (0: No, 1: Yes)	Definitions + Why/Why Not CoT (0: No, 1: Yes)
1	Understandability	F(2, 1621) = 4.961 **	0.1* (0.032)	0.113 (0.149)			
	Informativeness	F(2, 1621) = 3.668 *	0.09* (0.034)	0.124 (0.154)			
	Harm	F(2, 1621) = 3.876 *	0.024 (0.049)	-0.507* (0.224)			
2	Understandability	F(3, 1620) = 3.368 *	0.103* (0.033)	0.14 (0.161)	0.015 (0.034)		
	Informativeness	F(3, 1620) = 4.914 **	0.107* (0.034)	0.294 (0.166)	0.094* (0.035)		
	Persuasiveness	F(3, 1620) = 3.207 *	0.062 (0.036)	0.215 (0.173)	0.102* (0.036)		
	Harm	F(3, 1620) = 2.638 *	0.028 (0.05)	-0.47 (0.241)	0.02 (0.05)		
	Agreement	F(3, 1620) = 2.985 *	0.041 (0.031)	0.138 (0.149)	0.088* (0.031)		
3	Understandability	F(3, 1620) = 3.852 **	0.106* (0.033)	0.181 (0.158)		-0.042 (0.033)	
	Informativeness	F(3, 1620) = 7.053 **	0.108* (0.034)	0.328* (0.163)		-0.127* (0.034)	
	Transparency	F(3, 1620) = 3.408 *	0.054 (0.034)	0.438* (0.164)		-0.086* (0.034)	
	Persuasiveness	F(3, 1620) = 2.941 *	0.058 (0.035)	0.184 (0.171)		-0.095* (0.036)	
	Harm	F(3, 1620) = 3.932 **	0.01 (0.049)	-0.667* (0.237)		0.1* (0.05)	
	Agreement	F(3, 1620) = 7.267 **	0.046 (0.03)	0.204 (0.146)		-0.139* (0.031)	
4	Understandability	F(3, 1620) = 3.521 *	0.1* (0.033)	0.109 (0.149)			0.025 (0.031)
	Informativeness	F(3, 1620) = 2.76 *	0.089* (0.034)	0.118 (0.155)			0.031 (0.032)
	Harm	F(3, 1620) = 4.316 **	0.028 (0.049)	-0.487* (0.224)			-0.107* (0.047)

perceiving it as being too complex. In contrast, the "Definitions" prompting strategy showed advantages in several areas, suggesting that its simpler structure made it easier to for users to interpret and evaluate feedback. While the "Definitions + Why/Why Not CoT" strategy did not consistently deliver benefits, the C5 model was rated as the least offensive and most effective in the post-study survey. This strategy possibly encouraged the model to explore both sides of a decision, leading to more context-sensitive and intuitive feedback. It also aligned well with counterfactual explanations in explainable AI (XAI), where the model justifies its outputs by referencing plausible alternatives (Warren et al., 2024).

To address (**RQ3**) on whether the perceptions of LLM feedback associated with its performance on the focal task, results suggest a partial alignment between the two. Offensiveness was the only measure that was linearly associated with model performance. While the lowest-performing model (i.e., C6) did not consistently receive the lowest perceptual ratings, the highest-performing model (i.e., C5) was rated significantly higher in informativeness, transparency, and overall perceived effectiveness. This suggests that stronger task performance may correspond to more favorable user perceptions, particularly when models provide interpretable feedback, although the relationship is not strictly linear across all configurations. Even when a model's output is factually correct, human-centered evaluation metrics may not align with task performance. Therefore co-designing systems with stakeholders could further bridge this gap.

This study has some limitations: It focuses on a specific interview training task and one focal skill, thus the findings may not generalize to other domains or types of formative feedback. While MTurk provides access to a large participant pool, the population may not be fully representative of real-world job seekers. Given the rapid evolution of LLMs, model-specific findings from user studies may quickly become outdated. To mitigate this, alternative, scalable methods for approximating qualitative perceptions, e.g., Carton et al., (2020), could be valuable for early-stage evaluations without incurring high resource costs.

CONCLUSION

We evaluated the ability of LLMs to provide formative feedback on interviewees' responses in a job interview task. While LLM-generated feedback was perceived as somewhat useful, moderate ratings across key dimensions and higher perceptions of irrelevance and offnesiveness indicate that it may not yet meet the standards required for effective and trustworthy interview training. The choice of LLM and prompting strategy significantly affect user perceptions of feedback quality, with GPT generally rated more favorably than Gemini and simpler prompting strategies leading to more positively perceived feedback. Offensiveness was the only metric significantly associated with performance.

ACKNOWLEDGMENT

This research was funded by the National Science Foundation (#1763486).

REFERENCES

Aggarwal, P., Mahajani, G., Malasani, P. K., Jamadagni, V., Wendt, C. J., Nirjhar, E. H., & Chaspari, T. (2025). Investigating the Reasoning Abilities of Large Language Models for Understanding Spoken Language in Interpersonal Interactions. In Proc. Interspeech 2025 (pp. 4518–4522).

Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhart, P., Jones, H., Ochs, M., Pelachaud, C., Porayska-Pomsta, K., Rizzo, P., & Sabouret, N. (2013, November). The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. In International Conference on Advances in Computer Entertainment Technology (pp. 476–491). Springer International Publishing.

Carton, S., Rathore, A. and Tan, C. (2020) Evaluating and Characterizing Human Rationales. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 9294–9307).

Chen, Z., Chen, J., Singh, A., & Sra, M. (2023). XplainLLM: A Knowledge-Augmented Dataset for Reliable Grounded Explanations in LLMs. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 7578–7596).

De Angelis, M., & Miranda, S. (2023). A personalized feedback system to support teacher training. Research on Education and Media, 15(1), 30–39.

Dougrez-Lewis, J., Akhter, M. E., Ruggeri, F., Löbbers, S., He, Y., & Liakata, M. (2025, July). Assessing the Reasoning Capabilities of LLMs in the context of Evidence-based Claim Verification. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 20604–20628).

Gebhard, P., Schneeberger, T., André, E., Baur, T., Damian, I., Mehlmann, G., König, C., & Langer, M. (2018). Serious games for training social skills in job interviews. IEEE Transactions on Games, 11(4), 340–351.

Gurrapu, S., Kulkarni, A., Huang, L., Lourentzou, I., & Batarseh, F. A. (2023). Rationalization for explainable NLP: A survey. Frontiers in artificial intelligence, 6, 1225093.

He, M. and Garner, P. N. (2023). Can ChatGPT detect intent? Evaluating large language models for spoken language understanding?. Proc. Interspeech (pp. 1109–1113).

Hoque, M., Courgeon, M., Martin, J. C., Mutlu, B., & Picard, R. W. (2013, September). Mach: My automated conversation coach. In Proceedings of the 2013 ACM ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 697–706).

Joshi, B., Liu, Z., Ramnath, S., Chan, A., Tong, Z., Nie, S., Wang, Q., Choi, Y., & Ren, X. (2023). Are machine rationales (not) useful to humans? Measuring and improving human utility of free-text rationales. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 7103–7128).

Kalin, R., & Rayko, D. (2013). The social significance of speech in the job interview. In The social and psychological contexts of language (pp. 39–50). Psychology Press.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Proc. Advances in neural information processing systems, 35, 22199–22213.

Lang, G., Triantoro, T. and Sharp, J. H. (2024). Large Language Models as AI-Powered Educational Assistants: Comparing GPT-4 and Gemini for Writing Teaching Cases. Journal of Information Systems Education, 35(3), pp. 390–407.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. Personnel psychology, 67(1), 241–293.
Li, G., Chen, L. and Yu, K. (2023). How ChatGPT is Robust for Spoken Language Understanding?. In Proc Interspeech (pp. 2163–2167).
Lubos, S., Tran, T. N. T., Felfernig, A., Polat Erdeniz, S., & Le, V. M. (2024, June). LLM-generated explanations for recommender systems. In Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (pp. 276–285).
Lyu, Q., Apidianaki, M. and Callison-Burch, C. (2024). Towards faithful model explanation in NLP: A survey. Computational Linguistics, 50(2), pp. 657–723.
Naik, N. D., Abbott, E. F., Gas, B. L., Murphy, B. L., Farley, D. R., & Cook, D. A. (2018). Personalized video feedback improves suturing skills of incoming general surgery trainees. Surgery, 163(4), 921–926.
Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015, May). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG) (pp. 1–6).
Okoso, A., Otaki, K., Koide, S., & Baba, Y. (2025). Impact of tone-aware explanations in recommender systems. ACM Transactions on Recommender Systems, 3(4), 1–34.
Pennebaker, J. W.et al. (2015). The development and psychometric properties of LIWC2015.
Polonsky, W. H. and Fisher, L. (2015) When does personalized feedback make a difference? A narrative review of recent findings and their implications for promoting better diabetes self-care. Current diabetes reports, 15, 1–10.
Project VetTrain. Available at: https://sites.google.com/colorado.edu/vettrain/dataset (Accessed: 31 August 2025).
Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153–189.
Smith, A. B., Schieber, D. L. and Austin, T. L. (2020). Avoiding "Great job!": Self- and peer feedback as formative assessment for oral communication. Journal of the Academy of Business Education, 21, 100–123.
Warren, G., Byrne, R. M. J. and Keane, M. T. (2024). Categorical and continuous features in counterfactual explanations of AI systems. ACM Transactions on Interactive Intelligent Systems, 14(4), 1–37.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824–24837.
Xu, W.et al. (2023). Artificial intelligence in constructing personalized and accurate feedback systems for students. International Journal of Modeling, Simulation, and Scientific Computing, 14(1), 2341001.