Using LLMs to identify features of personal and professional skills in an open-response situational judgment test

Cole Walsh, Rodica Ivan, Muhammad Zafar Iqbal, and Colleen Robb
Acuity Insights Inc., Toronto, ON, Canada

Abstract

Academic programs are increasingly recognizing the importance of personal and professional skills and their critical role alongside technical expertise in preparing students for future success in diverse career paths. With this growing demand comes the need for scalable systems to measure, evaluate, and develop these skills. Situational Judgment Tests (SJTs) offer one potential avenue for measuring these skills in a standardized and reliable way, but open-response SJTs have traditionally relied on trained human raters for evaluation, presenting operational challenges to delivering SJTs at scale. Past attempts at developing NLP-based scoring systems for SJTs have fallen short due to issues with construct validity of these systems. In this article, we explore a novel approach to extracting construct-relevant features from SJT responses using large language models (LLMs). We use the Casper SJT to demonstrate the efficacy of this approach. This study sets the foundation for future developments in automated scoring for personal and professional skills.

1 Background

A longstanding challenge in academia is selecting qualified and professional candidates from a larger applicant pool for professional training programs. Decision makers have traditionally relied on measures of hard skills and cognitive ability to make these decisions (Eva et al., 2009), often relying on grade point average (GPA) and standardized tests such as the Scholastic Aptitude Test (SAT), Graduate Record Exam (GRE), Medical College Admission Test (MCAT), and Graduate Management Admission Test (GMAT). Personal and professional skills such as communication, teamwork, problem-solving, and critical thinking, although recognized as predictive of future success in education and industry (Heckman and Kautz, 2012), have been more difficult to measure for a number of reasons including lack of standardization and scalability (Patterson et al., 2016). Admissions committees have commonly used reference letters, personal essays, and interviews as a proxy for applicants' personal and professional skills, but these processes do not meet the psychometric standards that we would expect from tools used in high-stakes decision-making (Kuncel et al., 2014; Patterson et al., 2016). Additionally, as the adoption of generative AI spreads, there are increased concerns about the authenticity of reference letters and personal essays (Chen et al., 2024), further exacerbating the need for valid and reliable tools to measure personal and professional skills.

Recognizing the limitations of other admissions tools (i.e., reference letters, personal essays) (Patterson et al., 2016), higher education programs have been increasingly turning to a more reliable and standardized tool, Situational Judgment Tests (SJTs), to assess applicants' personal and professional skills as part of their admissions process (Webster et al., 2020; Nadmilail et al., 2023). Though they may be delivered in different formats, including fixed-response and open-response, SJTs generally involve simulated situations and questions designed to elicit how a respondent would likely react in the situation (Lievens, 2013). Fixed-response SJTs typically require respondents to select or rank possible actions based on their effectiveness in a given situation and show stronger relationships with measures of cognitive ability, rather than personal or professional skills (McDaniel et al., 2007). Open-response SJTs, on the other hand, are more conducive to measuring behavioral tendencies (i.e., how the respondent would likely react in the given situation) and tend to show stronger relationships with personal and professional skills relative to fixed-response SJTs (McDaniel et al., 2007).

Although open-response SJTs have proven effective in evaluating personal and professional skills in a standardized and reliable manner, there are

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 221–230
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

challenges in executing these types of assessments at scale. Open-response SJTs are primarily scored by human raters who require extensive training to become proficient at evaluating responses (Shipper et al., 2017). Additionally, performing this level of scoring at scale requires many trained human raters operating in parallel, which presents further operational barriers. These challenges are not unique to SJTs; developers of other open-response assessments have faced similar obstacles and overcome them with automated scoring systems such as Natural Language Processing (NLP) algorithms (Valenti et al., 2003). NLP-based scoring systems of this kind have been shown to achieve strong psychometric results in writing and language proficiency tests (Chodorow and Burstein, 2004; Ramineni et al., 2012; Cardwell et al., 2022), as well as short-answer tasks (Leacock and Chodorow, 2003).

While there is a growing literature on NLP-based scoring systems for open-response assessments, few studies have investigated their efficacy specifically for SJTs (Bulut et al., 2022; Walsh et al., 2022). One issue is that insights from other automated scoring systems may not be immediately transferable to SJTs given the difference in the measured construct: while other open-response assessments may focus on language proficiency or content-mastery, SJTs measure personal and professional skills (e.g., teamwork, problem solving, critical thinking) (Lievens and Motowidlo, 2016). These differences in the measured construct influence the kinds of features used as inputs to the scoring system. In particular, NLP-based scoring systems for writing and language proficiency typically use features related to coherence, grammar, and organization (Chodorow and Burstein, 2004; Ramineni et al., 2012; Cardwell et al., 2022), features which have no direct link with most constructs assessed by SJTs. Any valid NLP-based scoring system should exhibit construct relevance through the features used as inputs to said system (McCaffrey et al., 2022), making existing approaches to NLP-based scoring largely inapplicable to SJTs. Additionally, because open-response SJTs allow respondents to describe actions that they would take or have taken in the past, these assessments are designed to allow for complexity and response diversity, and thus there is generally no single correct answer (Dore et al., 2017). This characteristic of SJTs makes scoring responses based on "correctness" or similarity with other responses impractical as well.

2 Aims

In this study, we investigate the feasibility of identifying and extracting construct-relevant features from SJT responses. We build on the work of Iqbal *et al.*(Iqbal et al., 2025) who used a mixed-methods approach to identify nine construct-relevant features that influenced raters' evaluations of an open-response SJT. We probe whether and to what extent we can identify these features in SJT responses automatically using NLP-based approaches. Recognizing the complex and nuanced nature of these features, we decided to use Large Language Models (LLMs) for this task. Recent studies have demonstrated strong performance of LLMs for essay scoring (Lee et al., 2024) even in domains like divergent thinking (Organisciak et al., 2023), which, similar to SJTs, have been notoriously difficult to automatically score because of the complex nature of the construct. This work sets the stage for future endeavors to build an automated scoring system for open-response SJTs and similar assessment types.

3 Sample

We used data from the Casper SJT in this study. Casper is an open-response SJT that purports to measure respondents' personal and professional skills along the following competencies: collaboration, communication, empathy, ethics, fairness, motivation, problem solving, resilience, and self-awareness (Dore et al., 2017; Saxena et al., 2024). Casper presents respondents with a series of hypothetical scenarios that include either a text-based or video-based prompt. Text-based prompts include a short description of a situation while video-based prompts include trained actors enacting a scripted scenario. The respondents are then asked questions related to the prompt and given a fixed amount of time to respond. The data we used in this study included responses to both types of scenarios: video-based and word-based. An example of a situation depicted in a video-prompt scenario is given below:

> Chris and Jane are sitting together in a small meeting room. Their manager, Gary, enters to deliver a few brief comments before retreating to an adjoining work space. Chris gets up to approach Gary when he notices Gary focused on his phone instead of work. Jane tells Chris that she sees Gary on his phone very often and that overall he does not

> do a lot of work. Chris says it does not seem fair for someone like Gary, who is senior to them in the company, to do less work and be paid a lot more. In addition, Gary takes their hard work to present as his own, taking credit for their efforts.

Respondents were instructed that they were a coworker of Chris and Jane in this scenario and asked the following three questions:

1. How would you handle this situation with Gary, your manager? Explain your response.
2. Imagine that Gary was completing his work in a timely manner outside of normal hours, but still behaving inappropriately while in the office. Would this change your opinion? Why or why not?
3. Describe some serious issues that can occur when supervisors are not present for their team.

In addition to the different types of scenario prompts, Casper also includes two distinct response formats: respondents are either required to type their responses to the associated questions within the allocated time or record a video of themselves responding to each question. To simplify this study, we only examined scenarios with the typed-response format as analyzing video responses would have required either transcribing the responses or passing the video media itself to a multimodal AI model, fundamentally altering the procedure employed here. We leave the analysis of video responses to a future study.

Casper is a completely human-rated assessment where a respondent's responses to a scenario are rated together holistically on a 1-9 Likert scale. Trained human raters are provided with scoring guidelines which help them contextualize Casper competencies for each scenario and determine how effectively the responses addressed the questions asked. Additionally, Casper is norm-referenced, which means that raters are also instructed to score each response relative to the other responses they are reviewing for the same scenario. Raters do not, however, use an analytical rubric when rating in order to encourage diverse perspectives and interpretation during the rating process. This rating approach allows for responses exhibiting different characteristics to still receive high scores as the context and reasoning provided by the respondent are also taken into account.

The diversity in scenarios and responses makes Casper ideally suited to study underlying features of personal and professional skills in SJT responses. Previously, Iqbal *et al.* identified nine construct-relevant features that influenced Casper ratings. For the purposes of the current study, we selected seven of these features to investigate the applicability of LLMs for feature extraction, omitting two features related to competencies associated with specific scenarios. Given that different Casper competencies are probed in each scenario, we omitted these two features from our investigation as their analysis would have required prompting with more scenario and competency-specific information, which would have extended the complexity of this pilot study.

Table 1 shows the seven features we selected for this study. Iqbal *et al.* previously analyzed 27 responses from each of three Casper typed-response scenarios, ensuring a uniform distribution of responses at each scoring level (i.e., three responses for each 1-9 score assigned by human raters). Two researchers independently classified the construct-relevant features present in all responses using the levels noted in Table 1.

For the present study, we doubled the size of the dataset used, re-using the dataset collected by Iqbal *et al.* while adding 27 responses from each of three additional scenarios, which were again classified by the same two human raters as in the original study. Thus, the complete dataset in this study comprised 162 responses across six distinct Casper scenarios. We report agreement for the researchers' classifications of the features across all 162 responses in the last column of Table 1 for each of the seven features. We used Cohen's κ with quadratic weighting (after mapping features to a numeric scale) to measure agreement. In the case of binary features, the quadratic-weighted κ is identical to an unweighted κ.

4 Methods

We used LLMs as classifiers to replicate the work of the human raters in identifying construct-relevant features in Casper responses. Previous studies of LLMs for essay scoring have identified performance gains when LLMs are allowed to specifically evaluate one aspect of writing at a time (Lee et al., 2024). We applied a similar principle here by prompting LLMs to evaluate only one feature at a time for a given response.

We conducted two separate analyses. In our first study we compared several LLMs based on how well their classifications aligned with those of human raters for each feature. We used five state-of-the-art LLMs, listed in Table 2, including a mix of reasoning and non-reasoning, open and closed-source models.

For each LLM and feature, we generated classifications for all 162 Casper responses, then computed the κ between model classifications and each human rater's classifications. We then averaged the results to obtain one average κ for each LLM and feature. We used the same zero-shot prompt for all LLMs; we did not provide the LLMs with any examples within the prompts. Further, we wanted to compare how well each LLM performed using the same prompt without tailoring the prompt to work better with one LLM or another, so we provided only the necessary information to carry out the task within the prompt. Below is a minimal reproducible example of the system prompt we used:

> *You are a helpful assistant that analyzes users' responses to an ethical dilemma.*
>
> *The user was given the following prompt: "{context}".*
> *They were asked to respond to the following questions related to this prompt: "{questions}".*
>
> *Your task is to analyze {feature_description}.*
> *Return your response as a JSON object with the following keys:*
> *"decision": <{feature_levels}>*
> *"reasoning": <Reasoning for decision>*

where curly braces denote variable inputs in the prompt. {context} is the text prompt (for text-based scenarios) or the summary of the video (for video-based scenarios), {questions} is an enumerated list of the questions asked, {feature_description} is a description of the feature being evaluated (see Table 1), and {feature_levels} are the possible values that the LLM may assign.

We required the LLMs to return a "reasoning" for their decisions. For example, using the scenario and questions described above with the PERSP feature, o4-mini as the LLM, and the following response:

1. I'd invite Gary to a friendly chat, share what Jane and I feel, and listen to his pressures. Together we'd create small habits—visible credit tags and quick check-ins—so everyone's effort shines and Gary can still balance his senior duties.

2. If Gary finishes work later, I'd appreciate his flexibility, yet gently note that daytime presence and fair credit nurture team spirit. Leadership isn't only outcomes; it's modeling engagement and celebrating others while they're present.

3. When supervisors vanish, teammates can feel adrift: goals drift, growth pauses, mistakes linger, and trust erodes. Consistent guidance, recognition, and availability are anchors that keep performance, learning, and well-being afloat.

we get the following output:

> *"decision":"Thoughtfully considers and empathizes with multiple perspectives", "reasoning":"The user's responses address the concerns of the co-workers (feeling undervalued and needing recognition), invite Gary to share his pressures and perspective, and propose collaborative solutions that respect his senior role. They also weigh the team's needs for presence, credit, and guidance, demonstrating balanced empathy for both the manager and the team."*

We do not inspect "reasoning" outputs in this study, though such components can provide important validity evidence for LLM-based scoring systems (Casabianca et al., 2025), hence why we include this element in our prompting strategy.

In our second study, we investigated whether and to what extent we could improve LLM-human agreement by providing further details, including inclusion and exclusion criteria, for each level of a feature. Previous studies have indicated that this prompting strategy can effectively improve LLM performance in essay evaluation tasks (Lee et al., 2024). We worked exclusively with o4-mini for this analysis because it offered the best combination of throughput, cost, and performance that was ideally suited for this iterative pilot study.

5 Results

5.1 Comparison of LLMs with zero-shot prompt

Results are shown in Table 3, while Figure 1 shows the average κ agreement between each LLM and the two human raters on each feature using the zero-shot prompt. We find that Claude Sonnet 4 generally outperforms the other models, achieving the highest agreement with human raters on four out of seven features, while achieving the second highest agreement on two other features (JUST and CREAT). LACKINF was the lone feature where Claude Sonnet 4 did not rank among the top two LLMs, but even in this case the model achieved near human-level agreement ($\kappa_{\text{Claude Sonnet 4}} = 0.566$ compared to $\kappa_{\text{Humans}} = 0.640$).

o4-mini provides similar results to Claude Sonnet 4 in most cases, but notably struggles to identify responses that "provide insightful, novel, or creative arguments to address the questions" (CREAT). DeepSeek-R1, on the other hand, excels at identifying responses that fit this definition. GPT-4o mini is generally outclassed by the other models, but does reach super-human agreement in identifying responses that "state that they do not have enough information to make a decision" (LACKINF). Among all features explored in this study, LACKINF is the most likely to be identifiable through the use of particular words or phrases. For example, the string "gather more information" appears in seven responses. One human rater marked four of these responses as exhibiting the LACKINF feature while the other human rater marked all seven responses as exhibiting this feature. GPT-4o mini, meanwhile, classified six out of the seven responses as exhibiting LACKINF. Features such as this one that may be identifiable through keyword or semantic relationships likely see smaller benefits from using LLMs and, especially, more advanced reasoning models.

For four out of seven features, the top performing LLM achieved $\kappa > 0.4$. However, none of the LLMs approached human-level agreement on any feature outside of LACKINF. The disparities between LLM-human agreement and human-level agreement ranged from 0.209 to 0.352 ($\kappa_{\text{Humans}} - \kappa_{\text{LLM}}$). This result is not surprising given the sparse instructions provided to the LLMs in the zero-shot prompt. In the second part of this study, we investigate whether and to what extent we can close this gap via prompt engineering.

5.2 Improving LLM-Human Agreement

Disagreement between LLMs and human raters generally stems from lack of alignment on thresholds separating the levels of a feature. Table 4 shows the proportion of classifications made by the two human raters and o4-mini for each level of each feature. We can see that o4-mini is typically misaligned with the human raters in terms of how to separate the levels of a feature. For example, while human raters label a response as "fail[ing] to acknowledge or show sensitivity towards the legitimate concerns or feelings of one of the parties involved" (DISRES) only $5-6\%$ of the time, o4-mini classified 22.2% of responses as such. Similarly, o4-mini classified 59.9% of responses as having "Reasonable Justification", while classifying 0% and 0.6% of responses as displaying "No Justification" and "Clear and Compelling Justification", respectively. Human raters, meanwhile, provided more classifications at these extreme ends of the ordinal scale at the expense of labels in the middle of the scale. This result reflects an overall pattern we see across all non-binary features: o4-mini tended to provide more classifications in the middle of an ordinal scale than we observed with human raters. We used these results to motivate our prompt engineering strategy and further delineate feature levels.

We focused on six features for prompt engineering, omitting LACKINF where o4-mini was already achieving close to human-level performance. Results are displayed in the last column of Table 3 as well as Figure 2. We find that including additional details about the levels for a feature in the prompt effectively improves the LLM's agreement with humans. For most features we saw improvements of $0.08 < \Delta\kappa < 0.1$, but for DISRES we saw gains of $\Delta\kappa = 0.206$. With these prompts, o4-mini would've performed higher than all LLMs tested in the first experiment on all features except LACKINF (which we did not investigate improving) and CREAT, where o4-mini performs better, but is still outclassed by most other models.

6 Conclusions and Future Work

This study evaluated the feasibility of using LLMs to extract construct-relevant features from the Casper SJT. We found that reasoning models like OpenAI's o4-mini and Anthropic's Claude Sonnet 4 generally performed best at identifying these complex and nuanced constructs in responses, even

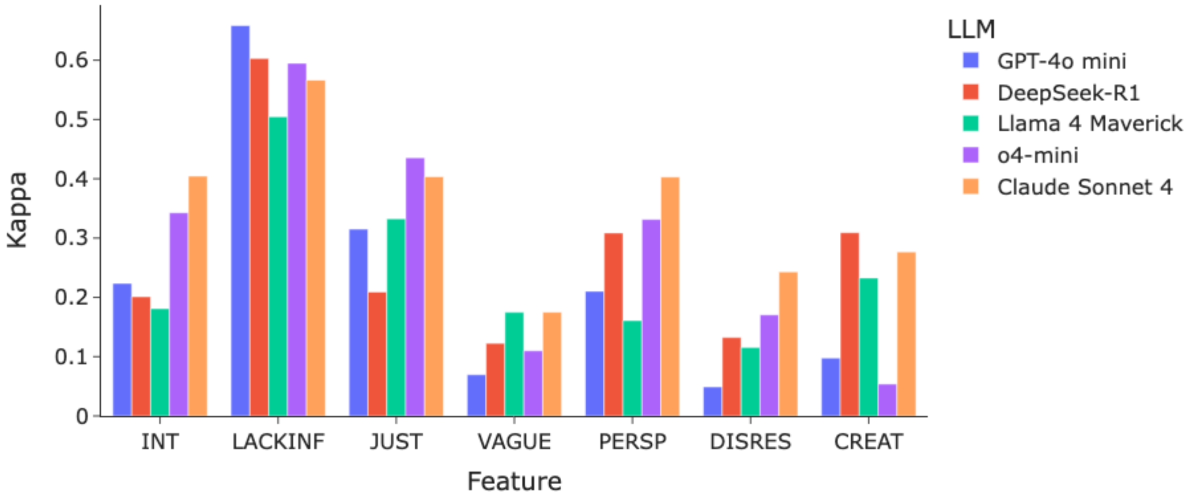

Figure 1: Average κ with human raters using zero-shot prompt.

with limited instructions. Additionally, we found that each LLM that we tested achieved the strongest performance on at least one feature. This result indicates that a future automated scoring solution using the same feature extraction method may be best served by using different LLMs for different features rather than forcing a single universal LLM. We could also consider using multiple LLMs for the same criteria and instituting a voting system resembling traditional machine learning ensemble methods to produce more accurate and reliable results (Dietterich, 2000). Overall, our results suggest a promising avenue for extracting construct-relevant features from SJTs and similar open-response assessments.

Prior to engaging in prompt engineering to improve performance, we had already reached close to human-level agreement in extracting one feature, whether a response "state[d] that they do not have enough information to make a decision." In this particular case, we hypothesize that features such as this one may be extractable by simpler models and methods such as keyword and semantic search.

For other features we fell well short of human-level agreement using zero-shot prompt classification, but we demonstrated that LLMs can be instructed to behave closer to expectations by giving further details about the levels for a feature. We found that providing these details had varying effects on our classification performance for different features, indicating that different features may be more susceptible to influence from this type of prompt engineering. Future work could explore other approaches to prompt engineering including few-shot prompting as well as fine-tuning to further improve performance.

We were also limited by small sample sizes in this study, owing to the effort and expertise required to annotate datasets such as these based on the features we explored. Future work will extend this study to explore a larger dataset from the Casper SJT as well as additional features. We plan to investigate the use of these features in an eventual automated scoring system for the Casper SJT. Such work would have important consequences, potentially extending the scalability and standardization of open-response assessments of personal and professional skills.

An automated scoring system based on the approach demonstrated here would also open avenues for future work in formative assessments by providing real time evaluation and feedback to respondents. We used a system prompt in this study that returned both a "decision" and "reasoning". We did not inspect the "reasoning"s in this study, but future work could use these "reasoning" fields to generate personalized and direct feedback for respondents. This method of extracting features from text could also be extended beyond assessments to other pieces of written text such as personal essays and reference letters.

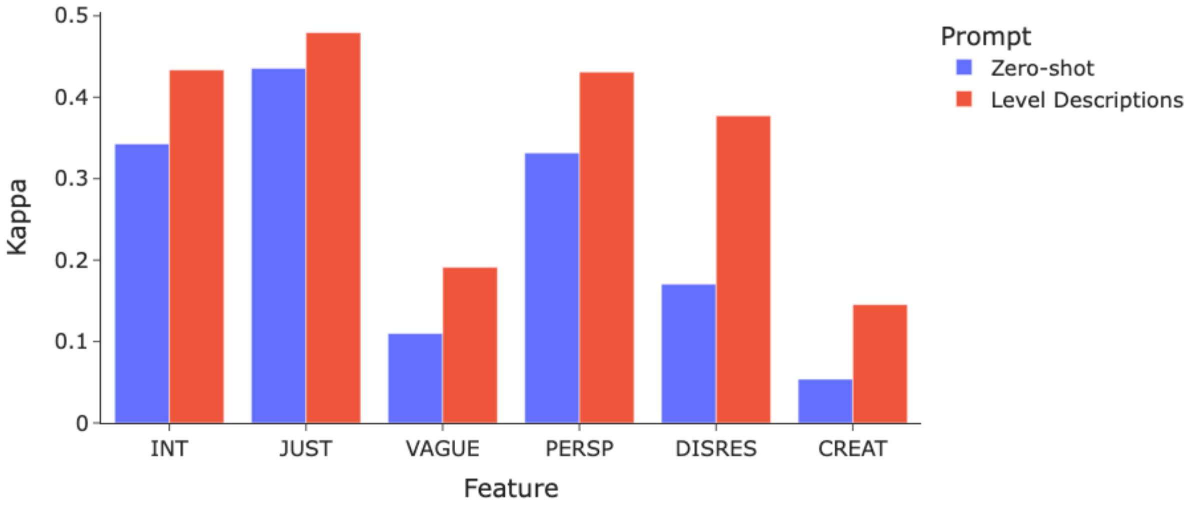

Figure 2: Average κ with human raters using o4-mini with zero-shot prompting and prompting with additional details for each level of a feature. Human-LLM agreement improves when providing additional level details in the prompt.

Acknowledgments

We would like to acknowledge Jillian Derby, Alexander MacIntosh, Josh Moskowitz, Gill Sitarenios, Susha Suresh, and Yiyu Xie for thoughtful discussions on the ideas presented in this manuscript.

References

Okan Bulut, Alexander MacIntosh, and Cole Walsh. 2022. Leveraging natural language processing for quality assurance of a situational judgement test. In *International Conference on Artificial Intelligence in Education*, pages 84–88. Springer.

Ramsey Cardwell, Geoffrey T LaFlair, and Burr Settles. 2022. Duolingo english test: technical manual. *Duolingo Research Report*.

Jodi M Casabianca, Daniel F McCaffrey, Matthew S Johnson, Naim Alper, and Vladimir Zubenko. 2025. Validity arguments for constructed response scoring using generative artificial intelligence applications. *arXiv preprint arXiv:2501.02334*.

Jeffrey Chen, Brendan K Tao, Shihyun Park, and Esta Bovill. 2024. Can chatgpt fool the match? artificial intelligence personal statements for plastic surgery residency applications: a comparative study. *Plastic Surgery*, page 22925503241264832.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Kelly L Dore, Harold I Reiter, Sharyn Kreuger, and Geoffrey R Norman. 2017. Casper, an online preinterview screen for personal/professional characteristics: prediction of national licensure scores. *Advances in Health Sciences Education*, 22:327–336.

Kevin W Eva, Harold I Reiter, Kien Trinh, Parveen Wasi, Jack Rosenfeld, and Geoffrey R Norman. 2009. Predictive validity of the multiple mini-interview for selecting medical trainees. *Medical education*, 43(8):767–775.

James J Heckman and Tim Kautz. 2012. Hard evidence on soft skills. *Labour economics*, 19(4):451–464.

Muhammad Zafar Iqbal, Rodica Ivan, Colleen Robb, and Jillian Derby. 2025. Evaluating factors that impact scoring an open response situational judgment test: a mixed methods approach. *Frontiers in Medicine*, 11:1525156.

Nathan R Kuncel, Rachael J Kochevar, and Deniz S Ones. 2014. A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment*, 22(1):101–107.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37:389–405.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213.

Filip Lievens. 2013. Adjusting medical school admission: assessing interpersonal skills using situational judgement tests. *Medical education*, 47(2):182–189.

Filip Lievens and Stephan J Motowidlo. 2016. Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, 9(1):3–22.

Daniel F McCaffrey, Jodi M Casabianca, Kathryn L Ricker-Pedley, René R Lawless, and Cathy Wendler. 2022. Best practices for constructed-response scoring. *ETS Research Report Series*, 2022(1):1–58.

Michael A McDaniel, Nathan S Hartman, Deborah L Whetzel, and W LEE GRUBB III. 2007. Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology*, 60(1):63–91.

Azad Iqram Nadmilail, Mohd Effendi Ewan Mohd Matore, Siti Mistima Maat, and Lynn Sheridan. 2023. Broad vs. narrow traits: a scoping review of measuring personality traits in teacher selection using the situational judgment test. *Frontiers in Psychology*, 14:1217321.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Fiona Patterson, Alec Knight, Jon Dowell, Sandra Nicholson, Fran Cousans, and Jennifer Cleland. 2016. How effective are selection methods in medical education? a systematic review. *Medical education*, 50(1):36–60.

Chaitanya Ramineni, Catherine S Trapani, David M Williamson, Tim Davey, and Brent Bridgeman. 2012. Evaluation of the e-rater® scoring engine for the gre® issue and argument prompts. *ETS Research Report Series*, 2012(1):i–106.

Anurag Saxena, Loni Desanghere, Kelly Dore, and Harold Reiter. 2024. Incorporating a situational judgement test in residency selections: clinical, educational and organizational outcomes. *BMC Medical Education*, 24(1):339.

Edward S Shipper, Laura M Mazer, Sylvia Bereknyei Merrell, Dana T Lin, James N Lau, and Marc L Melcher. 2017. Pilot evaluation of the computer-based assessment for sampling personal characteristics test. *Journal of Surgical Research*, 215:211–218.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330.

Cole Walsh, Alexander MacIntosh, Okan Bulut, and Jinnie Shin. 2022. What are we measuring?: A topic modeling framework to map professionalism aspects to responses in a situational judgment test. In *Companion Proceedings of the 12th International Conference on Learning Analytics Knowledge LAK22*, pages 39–41.

Elin S Webster, Lewis W Paton, Paul ES Crampton, and Paul A Tiffin. 2020. Situational judgement test validity for selection: A systematic review and meta-analysis. *Medical Education*, 54(10):888–902.

Key	Description	Levels	Cohen's κ (Humans)
INT	Grasps and addresses the complex social and emotional dynamics present in the ethical dilemma.	• Limited Interpretation • Adequate Interpretation • Excellent Interpretation	0.700
LACKINF	States that they do not have enough information to make a decision.	True/False	0.640
JUST	Justifies the course of action suggested.	• No Justification • Superficial Justification • Reasonable Justification • Clear and Compelling Justification	0.788
VAGUE	Vague or unclear.	True/False	0.356
PERSP	Considers the perspectives of the different parties involved in the dilemma.	• Considers one perspective • Briefly considers multiple perspectives • Thoughtfully considers multiple perspectives	0.722
DISRES	Fails to acknowledge or show sensitivity towards the legitimate concerns or feelings of one of the parties involved.	True/False	0.647
CREAT	Provides insightful, novel, or creative arguments to address the question.	True/False	0.510

Table 1: Features identified by Iqbal *et al.* as influencing Casper scores and used in the present study. Cohen's κ is reported between two independent human raters' classifications of these features across 162 Casper responses.

Table 2: LLMs explored in this study including their providers, whether they were reasoning models, and whether model weights were open or closed-source.

Name	Provider	Reasoning Model (Y/N)	Open/closed-source
GPT-4o-mini	OpenAI	N	Closed
DeepSeek-R1	DeepSeek	Y	Open
Lllama 4 Maverick	Meta	N	Open
o4-mini	OpenAI	Y	Closed
Claude Sonnet 4	Anthropic	Y	Closed

	Zero-shot					**Level Desc.**
Feature	**GPT-4o mini**	**DeepSeek-R1**	**Llama 4 Mav.**	**o4-mini**	**Sonnet 4**	**o4-mini**
INT	0.224	0.201	0.181	0.343	**0.404**	0.434
LACKINF	**0.658**	0.603	0.505	0.595	0.566	-
JUST	0.315	0.209	0.333	**0.436**	0.404	0.479
VAGUE	0.070	0.123	**0.175**	0.110	**0.175**	0.191
PERSP	0.210	0.309	0.161	0.332	**0.403**	0.431
DISRES	0.049	0.132	0.116	0.171	**0.243**	0.377
CREAT	0.098	**0.309**	0.233	0.054	0.277	0.145

Table 3: Average Cohen's κ agreement with human raters for each LLM on each feature using the zero-shot prompt. The last column shows the average κ for o4-mini after modifying the prompts to include level descriptions for each feature. We did not explore the LACKINF feature in this second experiment because we achieved close to human-level agreement with the zero-shot prompt.

		Proportion Selected		
Key	**Level**	**Human 1**	**Human 2**	**o4-mini**
INT	Limited Interpretation	0.435	0.377	0.327
	Adequate Interpretation	0.447	0.475	0.642
	Excellent Interpretation	0.118	0.148	0.031
LACKINF	False	0.944	0.889	0.864
	True	0.056	0.111	0.136
JUST	No Justification	0.062	0.056	0
	Superficial Justification	0.358	0.333	0.395
	Reasonable Justification	0.358	0.469	0.599
	Clear and Compelling Justification	0.222	0.142	0.006
VAGUE	False	0.790	0.883	0.568
	True	0.210	0.117	0.432
PERSP	Considers one perspective	0.302	0.407	0.149
	Briefly considers multiple perspectives	0.407	0.549	0.758
	Thoughtfully considers multiple perspectives	0.290	0.272	0.093
DISRES	False	0.938	0.951	0.778
	True	0.062	0.049	0.222
CREAT	False	0.833	0.796	0.994
	True	0.167	0.204	0.006

Table 4: Proportion of responses where each feature level was selected by human raters and o4-mini (with zero-shot prompt).