Towards Scalable Measurement of Durable Skills

Amir Globerson[a], Amy Keeling[a], Anisha Choudhury[a], Anna Iurchenko[a], Aviad Segal[d], Avinatan Hassidim[a], Ayça Çakmakli[a], Ben Gomes[a], Benn Witt[a], Cathy Cheung[a], Cristine Legare[c], Diana Akrong[a], Eliad Carmi[d], Elisabeth Bauer[a], Gal Elidan[a], Hadas Gelbart[d], Hairong Mu[a], Katherine Chou[a], Lev Borovoi[a], Nir Kerem[a], Niv Efron[a], Noa Kerrem Gilo[a], Preeti Singh[a], Rajvi Kapadia[a], Rena Levitt[a], Roni Rabin[a], Ronit Levavi Morad[a], Rotem Yulzary[a], Shashank Agarwal[a], Sophie Allweis[a], Tracey Lee-Joe[a], Tzvika Stein[a], Yael Bar Moshe[d], Yael Haramaty[a], Yaniv Carmel[a], Yishay Mor[a], Yoav Bar Sinai[a], Yoav Bergner[b], Yossi Matias[a] and Yuri Lev[a]
[a]Google Research, [b]New York University, [c]The University of Texas at Austin, [d]OpenMic

Durable skills, such as collaboration, creativity and critical thinking, are instrumental to success in the modern workforce. Yet, measuring these skills remains a persistent challenge. Moreover, because what is not measured is often not taught, these skills are often overlooked in mainstream educational curricula. Designing effective assessments for these skills necessitates balancing two often-conflicting requirements: ecological validity and psychometric rigor. On the one hand, the assessment environment should emulate natural interaction between humans, which is how these skills will be performed in the real world. On the other hand, it should be scalable, controllable and reproducible. Here we argue that Large Language Models (LLMs) can be used to better capture both of these aims. Concretely, we develop an AI-based framework where the subject converses with AI teammates in a way that resembles human-human interaction for authenticity, while also offering the psychometric control required for informative and robust assessment. Importantly, the AI participants not only act as teammates but also, in an "Executive LLM" setup, steer the conversation towards eliciting a high density of observable evidence for skill proficiency. We complement this with an AI evaluator that can be used to measure skill proficiency in such interactions. We evaluate our assessment protocol based on transcripts of interactions of human participants with our AI framework, for multiple durable skills. For the skill of creativity, we further demonstrate the efficacy of an autorater for evaluating complex tasks performed by high school students. Our analysis shows that the use of the Executive LLM significantly increases elicited evidence, compared to non-steered interactions. In addition, we show that LLM-automated scoring of conversations largely agrees with that of expert annotators. This research demonstrates the utility of orchestrated LLMs approaches for measuring complex social and cognitive constructs in a scalable and controllable manner.

Keywords: Future-ready skills, Durable skills, Scalable assessment, Generative AI assessment

1. Introduction

Success in the modern workplace requires not only technical knowledge and procedural fluency, but additionally a host of future-ready human skills, including collaboration, communication, critical thinking, and creativity [1, 2]. Often also referred to as 21st century skills [3–7], these constructs remain notoriously hard to measure [8–10]. Previous efforts to resolve these measurement challenges and align educational goals, assessment, and instruction have focused on technological innovation, for example through automated scoring of answers [11–13] and the collection and analysis of detailed process data [14, 15]. Progress in large language models (LLMs) has opened up a new technological frontier, which we pursue here. Our key idea is that LLMs can bridge the gap between unstructured student collaboration, which more closely emulates classroom practice, and standardized assessment, which, while artificial, attempts to isolate the behaviors needed for valid inference.

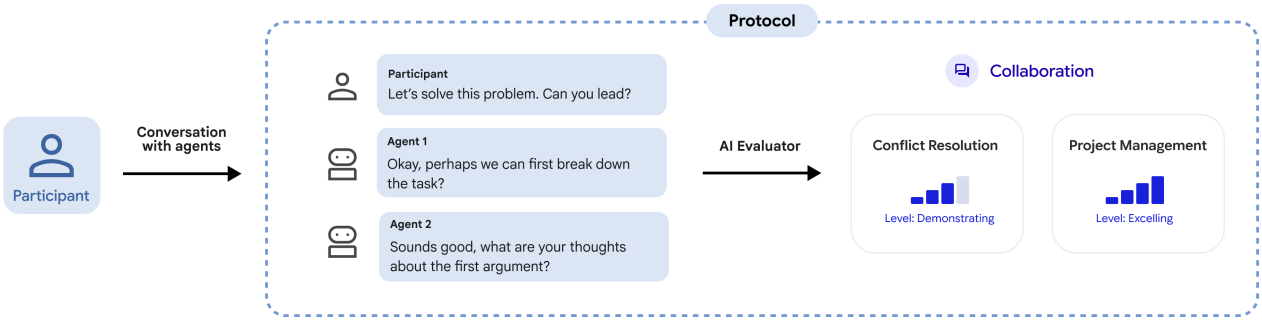

Figure 1 | An illustration of the AI-based assessment protocol. A human participant engages in conversation with AI-based teammates to solve a given task (e.g., collaboration in this example). After the conversation concludes, an AI Evaluator LLM analyzes the transcript and provides an assessment of skill levels along several dimensions (in this case Conflict Resolution and Project Management).

Group work is a key concept in education, and is viewed as beneficial to individual learning itself [16] and to the development of durable skills [17]. However, assessing the individual in such settings is especially challenging from a psychometric perspective due to the inherent interdependence between interacting members. Only a handful of large-scale efforts have addressed this challenge, the most notable being the Assessment and Teaching of 21st Century Skills project (ATC21S)[7] and the PISA 2015 CPS assessment [18–20]. While each program defined its own framework for the measurement construct, the differences in theoretical approaches were relatively minor. More profound were the different operationalizations of the construct. PISA employed an environment where the human subject interacted with scripted simulated teammates via multiple choice questions. In contrast, in the ATC21S project assessment tasks were carried out in human-human dyads, solving collaborative tasks by acting on objects in a digital environment (e.g., passing objects to one another) as well as communicating via a chat window. These approaches demonstrate two points on the spectrum between naturalistic and fully structured interactions, yet both are still far from true group interaction observed in classrooms.

Motivated by the above, we ask whether it is possible to develop an assessment approach that captures the open-ended nature of human-to-human interaction, while avoiding the variance of unstructured "in-the-wild" interactions. We argue that LLMs (e.g., [21, 22]) offer an effective approach to the problem. Prior to the availability of LLMs, incorporating artificial personas in a simulation for assessment required mostly hard-coded rules, thus constraining the simulated group work. This was intentionally the case for the human-agent tasks in large-scale assessments (e.g., PISA 2015 [20]).

Naturalistic conversational interactions are a native affordance of LLMs, opening the door to less structured, more authentic interactions. However, reduced structure often reduces the guarantee of observing the specific behaviors necessary for inference. Indeed, Sijtsma [23] observed that measurement is a compromise in the name of efficiency since the "long lasting observation of a person in real life until (s)he spontaneously exhibits the behavior of interest... would take too much time before enough evidence was collected." Here we address this challenge by introducing the Executive LLM, designed to explicitly drive the conversation towards eliciting evidence of specific skills and maximizing assessment accuracy, while keeping the conversation natural. Our approach can therefore be viewed as an adaptive test of complex behaviors.

We implement the Executive LLM within a scalable virtual assessment experience called Vantage, which resembles the flexible nature of human-human interactions while offering sufficient structure for robust assessment of future-ready skills. In Vantage, a human subject interacts with one or more AI-based teammates to carry out a joint task. The group tasks are designed to resemble authentic

classroom tasks, and a rubric is developed to evaluate the human participant in the virtual interactions along dimensions relevant to the skill of interest. The Executive LLM generates the responses for all the AI teammates in the conversation and is designed to steer the conversion toward maximal information and assessment accuracy. Finally, a separate LLM is used to assess the transcripts of the human-AI interactions. See Figure 1 for an illustration of the overall assessment protocol.

The above protocol was used for assessing the skills of collaboration, creativity and critical thinking, each in the context of group tasks appropriate for the elicitation of evidence for that skill. In order to validate the assessment approach, we focused on the complex skill of collaboration, and collected conversations of human participants interacting with Vantage as well as ratings of these conversations by expert pedagogical raters. We additionally analyzed simulated conversations for the skills of creativity and critical thinking.

We posited several key questions for our assessment approach. First, can the skills of participants be reliably scored by human raters based on transcripts from the conversations generated in Vantage? Second, can the Executive LLM improve evidence accumulation by steering these human-AI conversations? Third, can automated scoring using an AI Evaluator agree with human raters as consistently as human raters agree with each other? Finally, we ask whether such evaluation protocols can be refined by simulating human subjects. Our results provide positive evidence for all four questions, thus establishing the Executive LLM approach as a promising mechanism for assessing proficiency in complex durable skills. Finally, for the skill of creativity, we also show that for complex creativity tasks submitted by high-school students, a Gemini based AI Evaluator is an effective creativity assessor, on par with human expert raters.

2. A Scalable Assessment Protocol for Human Skills

We consider the setting where a human subject is to be evaluated on a skill S, and a multidimensional scoring rubric is given for that skill. In our setting, S will correspond to collaboration, creativity, and critical thinking. The outcome of the assessment is a rating of the subject on each of the dimensions of skill S according to the rubric. We propose to measure the skill level of the subject by having them converse with AI teammates. The human and AI teammates are given a task they need to solve together. We now briefly discuss the components involved in this setting, illustrated in Figure 1.

2.1. The Executive LLM

Perhaps the simplest approach to constructing a group of LLM teammates is to literally construct a group of LLMs, and allow these to interact with the human and with each other with no further constraints. However, as we show in the experimental evaluation, this “Independent Agents” setting does not result in sufficiently informative interactions and more control is needed to elicit interactions that can be robustly assessed for skill proficiency.

Instead of having separate LLMs model each of the AI teammates, we employ a single LLM which generates the responses for all of them. This Executive LLM has access to pedagogical rubrics that will also be used for assessment, and is prompted to generate text that will maximize the information and accuracy of the assessment protocol. Namely, it should create situations that will allow the subject to demonstrate their skill, so that their skill level can be quantitatively inferred. As an example, if “Conflict Resolution” is evaluated, the Executive LLM may have one of the AI teammates start a conflict and maintain it until evidence of conflict resolution skill has been observed. Figure 2 shows a sample excerpt from a conversation in the Executive LLM protocol, where the Executive LLM steers the conversation towards eliciting more information about the participant’s skill-related behavior.

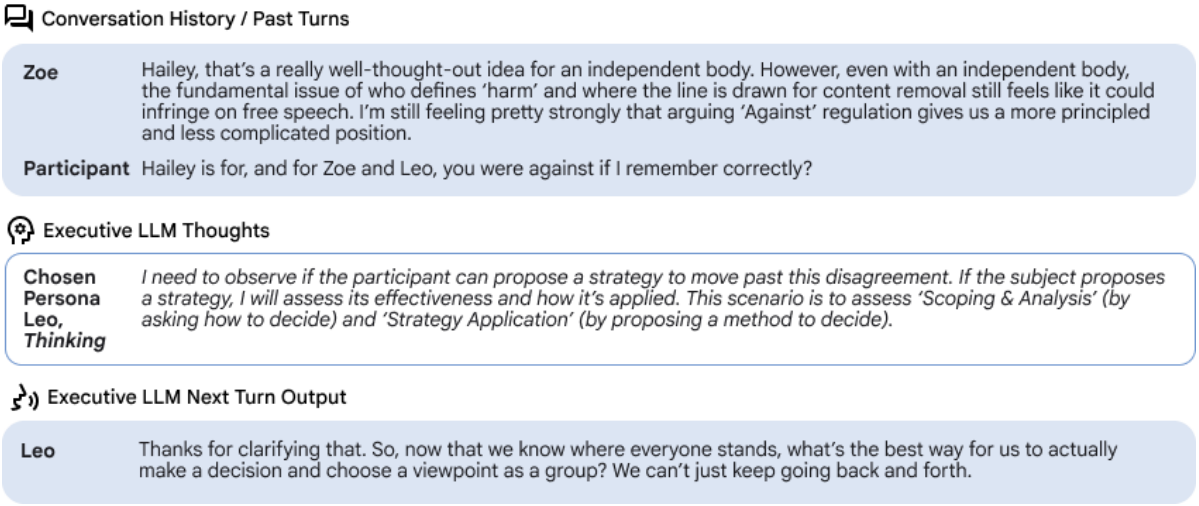

Figure 2 | An excerpt from a conversation between a human participant and the Executive LLM. The task is Debate and the Executive LLM focus is on conflict resolution. The Executive LLM thinks about the participant's response and chooses to respond as "Leo". Also shown is the internal reasoning of Executive LLM before generating the final response as "Leo".

2.2. Automatic Assessment and Feedback

Given a transcript of the interaction between the subject and the AI teammates for a task T and skill S and the corresponding evaluation multi-dimensional rubric (see below), our goal is to evaluate the level of the subject on each of the rubric dimensions. For each of the subject's turns in the conversation, we use an LLM (Gemini 3.0) to generate ratings at a turn level. The LLM was provided the conversation and prompted to output a level for each of the skill dimensions per turn. It also had the option to return NA, indicating no evidence for the given dimension in the turn. This rating was repeated 20 times per turn, and the final turn label was generated from these values, as follows. The final turn label was declared to be NA if at least one of the 20 predictions for the turn was NA. The final turn level rating (if not NA) was the most frequent level among the 20 returned. Then, to obtain a rating for the entire conversation, we took the turn outputs and trained a regression model (linear regression for the scores and logistic regression for the NA) based on human ratings. Model performance was evaluated using leave-one-out cross-validation.

Feedback is provided to the user in Vantage in a quantitative manner, along with qualitative evidence taken from the interaction with the AI teammates. An overall map gives a broad view of competency for all skills, with a break-down for different axes or sub-skills for each one. The user can further drill down and receive excerpts from the conversation which exemplify the numeric competency score. See Figure 3.

2.3. Skill Rubrics and Tasks

The Vantage system focuses on the core durable skills of collaboration, creativity and critical thinking. For each of those, we developed scoring rubrics, based on previous work. First, the relevant literature was reviewed in order to construct a conceptual model of the skill. From this, an initial rubric was derived. This rubric was used by human experts to assess sample conversations, and refined where agreement was low or the raters found it ambiguous. This approach is consistent with [24], who also found that LLMs are reliable for coding conversations when given a rubric that is derived from theory and then refined via expert use on sample data. Each rubric consisted of multiple dimensions, each

Figure 3 | The assessment and feedback provided to the user in Vantage. The skills map (top) shows quantitative competency of all skills practiced, along with a breakdown into the axes evaluated for each skill. Each can be expanded to also get qualitative feedback that explains the rating. The user can further drill down to a more detailed view (bottom) that provides excerpts from the conversation that substantiate the analysis.

with a score of 1-4, and an NA option, signifying there was insufficient information to decide on a score. See below for the conceptual basis of rubrics for each skill, and the appendix for the complete rubrics. The rubrics were provided as input to both the Executive LLM and the AI Evaluator.

For assessment and practice in Vantage, our pedagogical experts chose several themes for projects similar to those encountered in real classroom. For each theme, specific tasks for assessing and practicing the skills of collaboration, creativity and critical thinking were designed. See Figure 4 for a short description of the tasks for the creativity skill in Vantage. See the appendix for the conceptual basis of the tasks and rubrics for all skills.

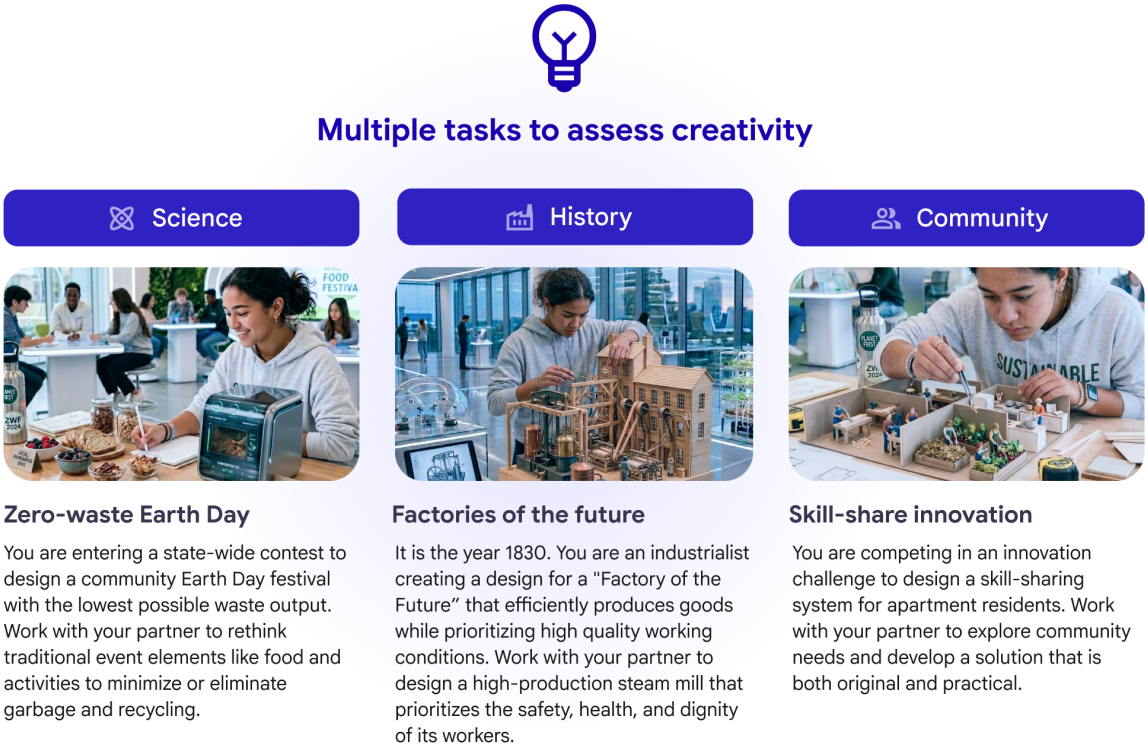

Figure 4 | Example tasks in Vantage for assessing and practicing the skill of creativity.

3. Vantage Experimental Setup

In our Vantage system a human subject interacts with virtual AI personas to complete a task aimed at eliciting sufficient information for skill assessment. We now describe the setup used for the evaluation of the Vantage skill assessment protocol. For evaluation, we focus mostly on the skill of collaboration where the user interacts with multiple AI teammates. We also present initial results for the creativity and critical thinking skills.

3.1. Environment

Participants who took part in the study were told they would be engaging in collaborative group activities with a set of AI teammates. Participants accessed a user interface that introduced them to their group of AI teammates and described the task they were to complete. Communication took place via a chat-like user interface which allowed interacting with the teammates. Each AI teammate had a name, and their contributions to the chat were delivered via both text and audio. Participants had the option to contribute to the discussion by either typing their responses or using a voice interface. Following the task description, they were requested to interact with the teammates for 30 minutes per conversation. Gemini 2.5 Pro [25] was used as model underlying the Executive LLM, that generated responses for all AI participants in the main body of experiments for the skill of collaboration. For the additional initial results for creativity and critical thinking in Section 4.2, Gemini 3 was used.

3.2. Participant recruitment

Participants ages 18-25 were recruited via the Prolific platform. They were English native speakers based in the United States. 188 participants were recruited, and each generated two conversations, resulting in a total of 373 conversations (three conversations were filtered due to technical issues).

3.3. Experiment design

In order to check the efficacy of the Executive LLM in steering skill-related behavior, we collected conversations driven by an Executive LLM, and also with Independent Agents. Specifically, we considered an Executive LLM that is prompted to focus on assessment of one of the sub-dimensions of collaboration: Conflict Resolution or Project Management. For each of the tasks, the protocol was randomized to be either an Executive LLM focusing on Conflict Resolution or on Project Management or to use Independent Agents. We also tested whether telling the subject to focus on the skill had an effect. Namely, before the conversation started we added the text: "In the following task, you are asked to pay particular attention to your role in [skill]." where skill was either "Conflict Resolution" or "Project Management". We randomly assigned subjects to receive either no focus instructions (50%) or instructions focusing on one of the two skills (25% each).

3.4. Human rating of conversations

Two pedagogical raters from New York University were recruited to annotate the conversations. They were provided with a scoring rubric for conflict resolution (CR) and for project management (PM). Raters were asked to consider each human turn of the conversation and, first, decide whether it should count as evidence of CR, or PM, or not. If not, they selected NA from a pull-down menu. If so, they selected a particular component of CR or PM from the rubric and provided a score on a scale of 1-4. In addition to the turn by turn rating, raters were also asked to provide a holistic score for the conversation. This holistic score could again be NA, if there was not enough evidence to decide, or a number between 1-4.

3.5. Simulated Conversations

Since evaluation with human participants is costly, it is worthwhile to have a sandbox that can simulate human participants, towards developing the protocol before deployment. Towards this end, Gemini was used to simulate the human participant in the conversation. For the recovery evaluation described in Section 4.1.5, Gemini was prompted to behave like a student evaluated on their collaboration skill, and was also instructed to behave according to a specific level of one of the rubric dimensions (e.g., level 3 of conflict resolution). This skill level was later compared to the level returned by the AI Evaluator. Each conversation contained 50 turns, and each level was repeated 100 times. We also used simulation to assess evidence level (e.g., Figures 10 and 9), and in these cases the simulated subject was not provided with a skill level.

4. Results

We start by assessing the validity of our assessment approach, evaluating the quality of the AI Evaluator as well as the efficacy of the Executive LLM, using both interactions with human participants as well as simulated subjects. These results are reported for the collaboration skill in Section 4.1. We then provide additional results on other skills in Section 4.2. Finally, we demonstrate the efficacy of our assessment autorater in a real student setting for a complex creativity task in Section 4.3.

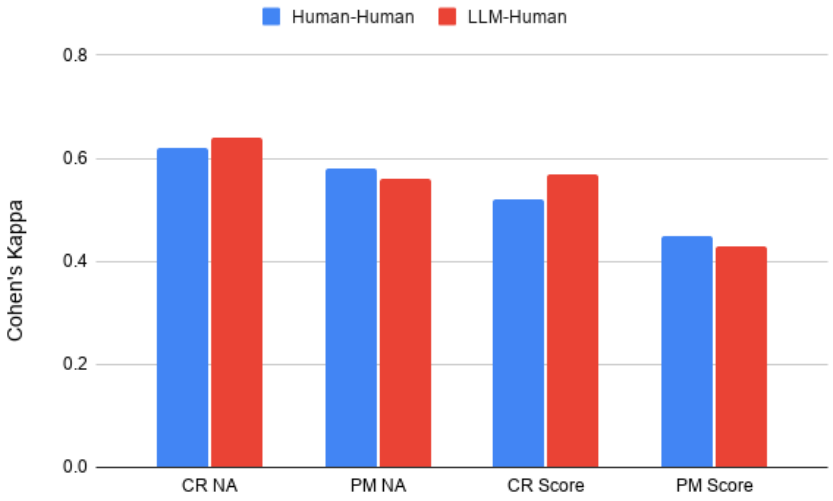

Figure 5 | Conversation level agreement using Cohen's Kappa (κ) for binary outcomes (NA or not), and quadratically weighted Kappa for scores, for conflict resolution (CR) and project management (PM). Compared are the inter-expert agreement (blue) and the LLM-expert agreement (red).

4.1. Evaluation of the Vantage assessment protocol for Collaboration

We carry out a comprehensive evaluation of our approach in the context of the collaboration skill, where the group setting is the most challenging and involves four members – the human participant and three AI team members.

4.1.1. LLM rating is on par with expert human raters

Roughly half of the conversations received a single human expert rating, and half of those also received an additional human expert rating. This allowed for a measure of inter-rater agreement between the two human experts and between the LLM AI Evaluator and the human experts. Agreement statistics using Cohen's Kappa [26, 27] for conflict resolution (CR) and project management (PM) are shown in Figure 5. These include score-or-no-score (i.e., NA) decisions, as well as finally numerical score agreements (for cases where both raters provided a numerical score).

Coding conversations for conflict resolution and project management proved to be challenging, even after several rounds of rater calibration. Agreement between the human expert raters was moderate, with Kappa in the range of 0.45-0.64 [28]. Agreement between the LLM and the human experts was similar to that between the two experts. This suggests that automatic rating of conversations is a scalable alternative to human rating in our assessment setup. Having established that the LLM-based autoraters provide similar results to those of human raters, below we use only autoraters, which can be run at a larger scale, to substantiate the validity of the assessment protocol. Results with human ratings are qualitatively similar to those of the autoraters.

4.1.2. Executive LLMs Elicit Skill-Related Information

The goal of the assessment task design is to elicit evidence of the skill of interest. A conversation-based group task *can* create opportunities for the subject to demonstrate conflict resolution (or project management) skills. But such evidence need not necessarily arise, for example if the group happens to work well and with no friction. We hypothesized that conversations using the Executive LLM protocol

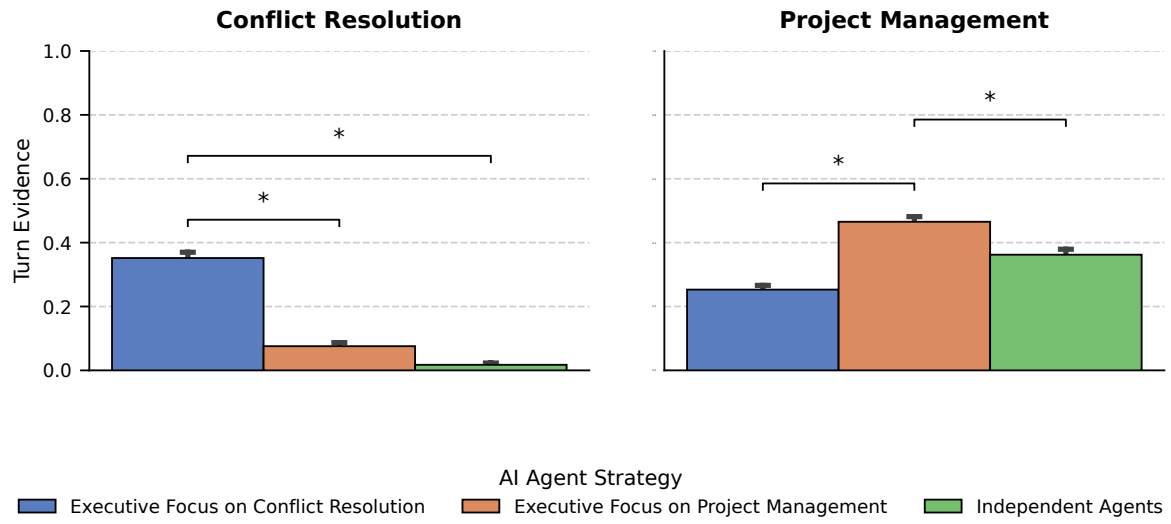

Figure 6 | Turn-level skill evidence. The fraction of participant turns rated as demonstrating a given skill for conflict resolution and project management. Results are shown for two versions of an Executive LLM, each focusing on a different skill, and the Independent Agents protocol. Starred brackets denote a statistically significant difference ($p \leq 0.05$) using the Fisher exact test.

would do a better job of eliciting evidence of the target collaboration skills. If so, this effect would be observable both at the turn-level and at the conversation-level, though not necessarily in the same quantity. The measure of evidence elicitation at both levels was simply calculated as the fraction of turns or conversations that were coded as evidence for the skill as opposed to as NA.

Figure 6 (turn-level) and Figure 7 (conversation-level) show evidence levels for Conflict Resolution (CR) and Project Management (PM) in each of three different AI agent protocols: a CR Executive LLM, a PM Executive LLM and Independent Agents. First, it can be seen that an Executive LLM focused on a skill always elicits significantly more evidence for that skill than the Independent Agents. Second, there is a desirable cross-over effect between the two versions of the Executive LLM when matched or mismatched to the measured skill. That is, steering the conversation towards CR increases evidence of CR but reduces evidence of PM and vice versa. This results in a significant difference in three of the four figures, except for measuring PM at the conversation level (Figure 7, Right), where similar information levels are obtained for both Executive LLM versions. Finally, the conversation-level information rate is quite high, at 92.4% for PM and 85% for CR, when the skill-matched Executive LLM was used. Taken together, these results indicate that use of Executive LLM extracts more informative interactions with respect to the skill that the Executive LLM is focused on.

It is also interesting to note that evidence rates are generally higher for Project Management than for Conflict Resolution, for almost all cases (except Executive Focus on PM at the turn level). This agrees with the intuition that project management behaviors are more abundant, and also require less steering than conflict resolution.

4.1.3. The Effect of Task Topic - Science vs Debate

The experimental setup involved two distinct tasks that subjects solved with the AI teammates: science and debate. To determine whether the type of task had an effect on conversation informativeness, independent of the influence of the Executive LLM, a logistic regression analysis was conducted. Separate models were fitted to predict the binary informativeness scores for both skills. To evaluate

Figure 7 | Conversation-level skill evidence for the collaboration skill. The fraction of conversations that were rated as having sufficient information to provide a skill-level rating (i.e., not rated as NA). Results shown for the Conflict Resolution (Left) and Project Management (Right) sub-skills. Results are shown for two versions of an Executive LLM, each focusing on a different skill, and the Independent Agents protocol. Starred brackets denote a statistically significant difference ($p \leq 0.05$) using the Fisher exact test.

the impact of the task type, the coefficient for the task variable in each model was tested against the null hypothesis that it was equal to zero. The results revealed no statistically significant difference from zero for either metric (p = 0.18 for Conflict Resolution informativeness; p = 0.9 for Project Management informativeness). Consequently, the analysis establishes that the assigned task does not significantly alter the informativeness of the conversation. This demonstrates the potential for using the the same framework on varied tasks and subject areas.

4.1.4. Is Telling the Subject Sufficient Without Steering?

The Executive LLM is meant to elicit skill-related behavior from the subject. We checked if such elicitation can be obtained by simply asking the subject to focus on a skill. To test this, subjects were randomly selected to either receive focus instructions or not. The focus instructions told the subject to focus on a particular skill. When the protocol was Executive LLM that skill was the same skill that the Executive LLM was focused on. Otherwise it was randomly chosen. To test the statistical effect of the focus on conversation informativeness, we used Fisher's exact test to check whether setting subject focus to a skill improved conversation information levels vs the no focus setting. Results showed that subject focus had no significant effect, both in the Independent Agents and Executive LLM settings and for both skills (all at $p > 0.6$). This suggests that the high information rates observed for Independent Agents cannot be achieved via subject focus instructions alone.

4.1.5. Executive LLMs Improve Skill Assessment - Recovery of "Known" Skill Proficiency of Simulated Subjects

Simulation studies (sometimes called Monte Carlo studies) in the psychometric literature are commonly used to evaluate new methods of inference [29]. Using item response theory, for example, it is straightforward to generate response probabilities from a latent trait and its distribution. Recovering

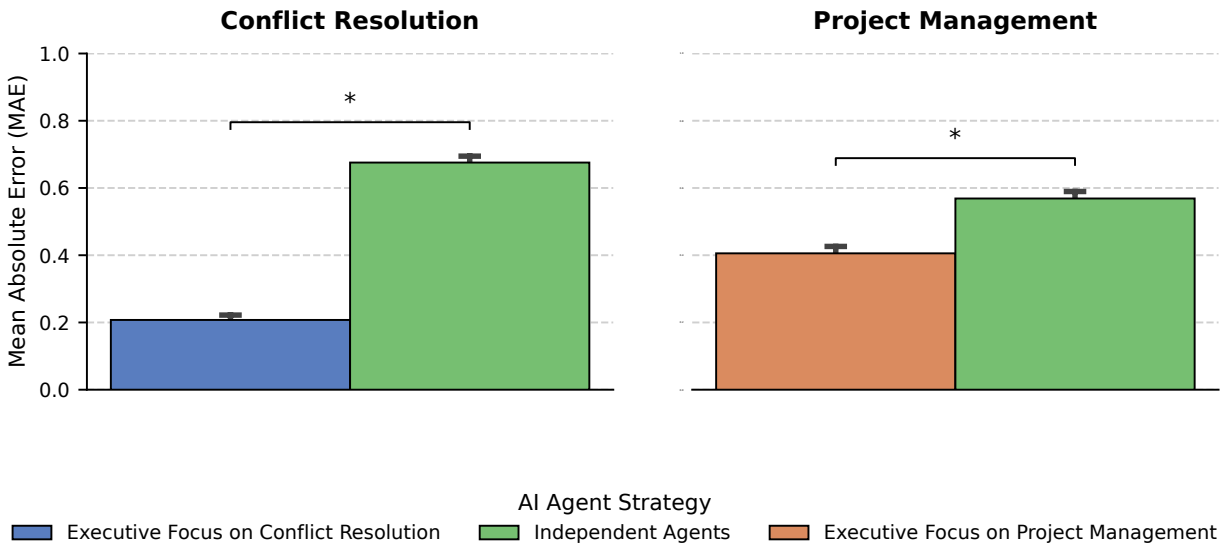

Figure 8 | Recovery error for different protocols. An LLM is prompted to simulate a subject at a given skill level L. The simulated subject then goes through the assessment procedure and the conversation is mapped to a skill level $\hat{L}$ using an autorater. The reported MAE is the mean absolute difference $|\hat{L} - L|$. Results are reported for both the conflict resolution and project management skills, comparing the Independent Agents and the corresponding Executive LLM protocol for each. Starred brackets denote a statistically significant difference using the Student's t-test.

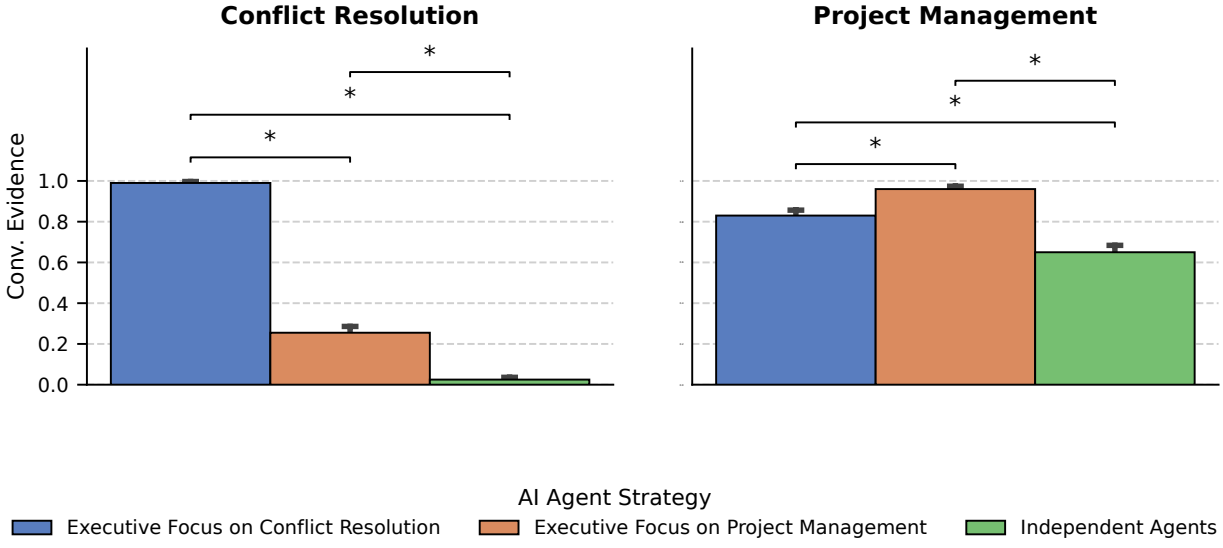

Figure 9 | Conversation-level skill evidence for simulated conversations. The setting is as in Figure 7, but the subject is simulated using an LLM. Starred brackets denote a statistically significant difference using the Fisher exact test.

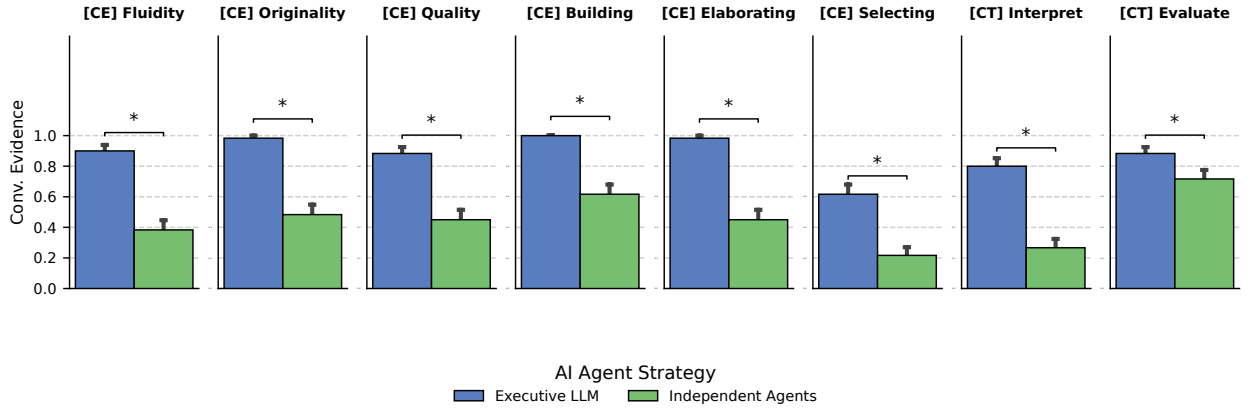

Figure 10 | Conversation-level skill evidence for the different axes of Creativity (CE) and Critical Thinking (CT), for simulated conversations. The setting is as in Figure 7 but for each skill (CE and CT) we use one Executive LLM focused on all dimensions of this skill, and the evidence rate for that Executive LLM is reported. For Critical Thinking the full names of the dimensions are "Interpret and Analyze" and "Evaluate and Judge", and for Creativity "Building" is short for "Building on Ideas".

the parameters used to generate the data, especially when complexity arises due to mixtures of populations, data censorship, or other confounds, is one step in validating the inferential capabilities of the model. Less commonly, process data such as sequences of steps or actions may also be generated from a sequence model, such as a Markov model (e.g., [30]), and then used again to test the inference methodology. Generative AI and LLMs constitute a relatively recent, non-parametric approach to both generating simulated data and inferring traits.

In the present study, we used the conflict resolution and project management rubrics as prompt inputs to a simulated collaborator. This was effectively another agent but one who is "external" to the task, as opposed to the task team Independent Agents and/or Executive LLM. We start by defining the "ground-truth" conflict resolution or project management skill level of the subject, each on a scale of 1-4, and then run the task with the simulated subject 100 times for each version of the protocol. We then rate each conversation using the same autoraters for CR and PM as used for the real data.

Figure 8 reports recovery results for the different assessment protocols. Recovery error is measured as the mean absolute difference (MAE) between the "true" and inferred skill scores. It can be seen that Executive LLM results in overall improved recovery rates, relative to Independent Agents. In addition, we can use the simulated conversations to calculate the same evidence rate metrics calculated on human conversations. Figure 9 reports information level in the simulated conversations. Comparison of Figures 7 and 9 shows that results are qualitatively similar between the human and simulated conversations. As might be expected, fully simulated conversations are a bit too ideal in that with the corresponding Executive LLM, the conversation evidence rate approaches 100%. Taken together, the results suggest that rubric-based LLM simulation can be a sandbox for developing improved assessment protocols, reducing some of the costly collection of data with human subjects.

4.2. Evidence Rates for Critical Thinking and Creativity

Above we presented detailed results for the skill of Collaboration, where we collected human ratings, and compared those to LLM-based ratings. For critical thinking and creativity, collection of human ratings is ongoing, and results will be shared at a later time. Here we present initial results for these skills, on conversations where the subject is simulated (as in Section 4.1.5 for the skill of collaboration).

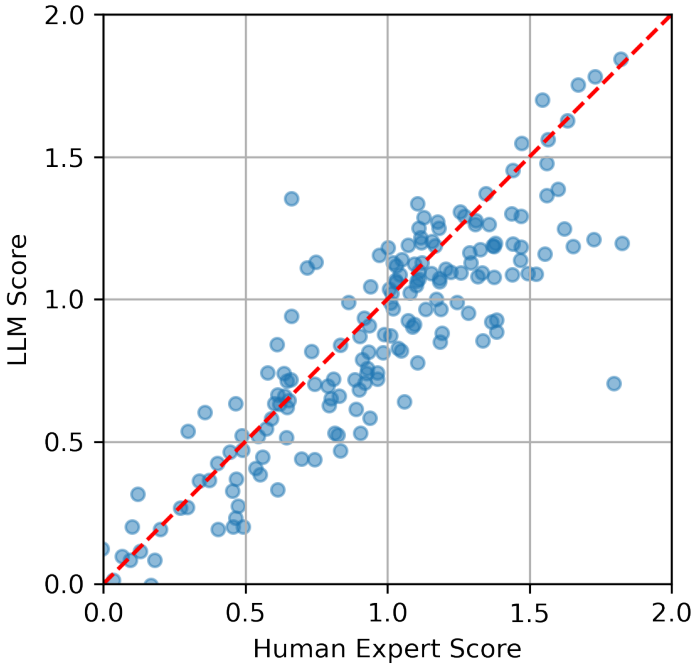

Figure 11 | Scatter plot of the autorater scores vs. the human expert scores for held out OpenMic's creativity task submissions. Pearson's correlation between the two is 0.88.

Figure 10 shows conversation-level evidence rates for the different dimensions of the creativity and critical thinking skills, when using either Executive LLM or Independent Agents as the AI participants.[1] As with collaboration, it can be seen that the use of the Executive LLM achieves conversation evidence rates that are higher that Independent Agents. We note that these ratings still need to be compared to human ratings, but this does suggest that the Executive LLM approach is also effective for the creativity and critical thinking skills.

4.3. Creativity Assessment in The Real World

In a partnership between Google and OpenMic, a startup developing AI-powered tools for assessing durable skills, we set out to evaluate the efficacy of our Gemini-based creativity autorater to score complex multimedia creativity tasks given to real students. Briefly, in a lab study, 280 students were given a short story and were then asked to design a news segment based on the story. This included multiple specific creative tasks such as coming up with character interview questions. The entire experience including the choice of story and specific tasks was designed by OpenMic's creativity experts. These experts also designed detailed rubrics for scoring each item on a scale of 0-2.

The completed tasks were graded by both trained experts and a Gemini-based creativity autorater which took as input the same rubrics used for expert grading. 100 submissions were used to modify and improve both the Gemini prompt and the expert pedagogical rubrics. The other 180 submissions were then used to evaluate the accuracy of the autorater relative to the human expert raters.

At the specific item $0-2$ scores, the agreement between the scores of the autorater and those of the human expert raters as measured using Cohen's Kappa was 0.66, corresponding to good agreement [28]. More importantly than the per item agreement, a student's proficiency is not measured based

[1]The NA conversation-level rating here was obtained by instructing Gemini to output a minimum and maximum score to assign the conversation on a given dimension, and return NA if these were not identical.

on a single item and thus we compare the total grade or sum of the scores of all the items in the submission. A comparison of the autorater overall scores and those of the human experts reveals an extremely high Pearson's correlation of 0.88. Figure 11 also shows qualitatively that our automated assessment of creativity for this complex task aligns strongly with those of human experts.

5. Discussion

We have argued that large language models have the potential to transform the assessment of complex durable skills. In the case of group work, they help overcome a number of fundamental psychometric challenges regarding the reliability, comparability, scalability, and ecological validity of such assessments. Until now, these challenges have been mostly addressed by highly scripted interactions with AI teammates (e.g., PISA 2015) or highly structured human-human interactions (ATC21S). The novel introduction of the Executive LLM allows standardization of the collaboration experience, without overly scripting the interaction itself.

Group work as defined here is still a human-to-human interaction (although human-agent interaction is rapidly becoming an interesting domain in itself). It follows that the most ecologically valid [31] form of group work assessment would be to put human subjects in groups for collaborative tasks. However, not only is this logistically challenging to scale, but the other team members (relative to any one assessed subject) introduce uncontrolled variance, hence lowering the reliability and comparability of the assessment procedure. Moreover, even if a subject is assessed several times in different groups, in order to counterbalance the variance from other group members, evidence of particular durable skills may or may not manifest naturally. E.g., how can we evaluate conflict resolution if team members all happen to agree?

One way to reduce this variability is by having human actors or "simulators" who interact with a subject, following a semi-scripted task. For example, simulation has long been used in medical education [32, 33] since the 1960s, and has been shown to be a reliable and effective tool for training and assessment of both core clinical skills and "soft skills" such as collaboration and communication [34]. Inspired by this tradition, simulations have expanded into other domains with a clinical character, such as teacher education [35], as means for training and assessing complex skills. However, scripting tasks with human subjects involves a high degree of tailoring, and is unlikely to be feasible outside of high-stakes contexts. Indeed, since the introduction of LLMs, their use has been explored for simulating human behavior in interaction. Some example applications are training of mental health professionals [36], training teachers [37], and medical education [38].

Replacing human subjects with AI teammates as group members inevitably trades some authenticity in exchange for comparability and reliability. AI teammates powered by LLMs are, at least, much more naturalistic than the rule-based teammates of the past. However, with that freedom comes less control of the conversation. If the conversation is completely unconstrained, then it likely will not be very efficient in providing the evidence of interest about the subject. The Executive LLM presented here addresses this problem by steering the conversation dynamically.

The Executive LLM approach is analogous to a computerized adaptive test (CAT) designed to increase or decrease the difficulty of test items so as to maximize information about the test-taker [39]. Whereas traditional CAT varies only the difficulty level of items in a narrow domain (e.g., mathematics or verbal skills), the Executive LLM can vary the flow of a conversation in a group task. It does so by modifying the AI participants collectively so as to heighten (or soften) conflicts, create the need for greater project management, or elicit additional creative ideas from the participant.

Our results on skill evidence (Figures 6 and 7) demonstrate that an Executive LLM focused

on a given skill can increase the evidence level for the skill of collaboration. This effect is more pronounced in the case of sub-skill of conflict resolution, where the Executive LLM can introduce conflicts explicitly, thus inducing the assessed subject to demonstrate strategies for resolving conflicts. For project management, the Executive LLM still elicits significantly more evidence as compared to Independent Agents, but differences are smaller. This can be explained by the fact that project management behavior arises naturally in conversation, and thus benefits less from steering by the Executive LLM.

Perhaps the most compelling validity evidence is Criterion Validity [40], measuring correspondence between the test results and external measures of performance (either concurrent or predictive). The most reliable version of these would be performance-related metrics such as manager reports, or teacher reports over semesters. However, collecting these is largely impractical due to time and privacy concerns. Thus, we instead propose a measure based on subject simulation, where the test is applied to a simulated subject instead of a human one. Recent results have indeed shown that simulated users can be used to model responses of human subjects [41–43]. The simulation results we present in Figure 9 are in qualitative agreement with those collected from conversation with human subjects in Figure 7. Furthermore, simulation allows comparison to a "ground truth" skill-level, and recovery results in Figure 8 support the efficacy of Executive LLM as compared to Independent Agents. These results suggest that simulation could be an effective approach towards designing assessment protocols for complex constructs such as collaboration, with improved validity. Finally, we also provided preliminary simulation results for the skills of creativity and critical thinking. The results suggest that the Executive LLM approach is also effective for these skills but further validation with human subjects is needed. We leave this to future work.

Our results for collaboration also demonstrate that LLMs can be used to score conversations according to a rubric, and their agreement with human raters is similar to inter-rater agreement between humans. These results join a growing body of work showing that LLMs can be used to considerably scale the assessment process [44–46]. However, our approach goes beyond rating human-to-human conversations using LLMs, and uses AI for creating conversations with human subjects, in an adaptive manner that provides skill-related information. Finally, we also demonstrated the efficacy of automated assessment for the skill of creativity in the context of a complex creativity task with real students in a lab setting.

This work focused mostly on developing an AI-based protocol and tool for assessing skills in a group setting. However, AI has great potential for *improving* these skills as well. As a simple example, the assessment environment can also be used for continual practice within a low risk sandbox environment, alongside standard group work in the classroom. We leave this possibility for future work. We also recognize that human skills are culturally situated, and will therefore also focus future work on exploring performance across diverse cultural settings and languages to ensure our technology is inclusive and equitable.

References

[1] Jeremy Burrus, Teresa Jackson, Nuo Xi, and Jonathan Steinberg. Identifying the most important 21st century workforce competencies: An analysis of the occupational information network (o* net). *ETS Research Report Series*, 2013(2):i–55, 2013.

[2] Linda Darling-Hammond, Joan Herman, James Pellegrino, Jamal Abedi, J Lawrence Aber, Eva Baker, Randy Bennett, Edmund Gordon, Edward Haertel, Kenji Hakuta, et al. Criteria for high-quality assessment. *Stanford Center for Opportunity Policy in Education*, 2:171–192, 2013.

[3] Bernie Trilling and Charles Fadel. *21st century skills: Learning for life in our times*. John Wiley & Sons, 2009.

[4] Joke Voogt and Natalie Pareja Roblin. A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of curriculum studies*, 44(3):299–321, 2012.

[5] Margaret L Hilton and James W Pellegrino. *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press, 2012.

[6] Foster Natalie and Piacentini Mario. *Innovating assessments to measure and support complex skills*. OECD publishing, 2023.

[7] Patrick Griffin, Barry McGaw, and Esther Care. *Assessment and teaching of 21st century skills*, volume 10. Springer, 2012.

[8] Brian M Stecher and Laura S Hamilton. *Measuring Hard-to-Measure Student Competencies: A Research and Development Plan. Research Report.* ERIC, 2014.

[9] Lei Liu, Jiangang Hao, Alina A von Davier, Patrick Kyllonen, and Juan-Diego Zapata-Rivera. A tough nut to crack: Measuring collaborative problem solving. In *Handbook of research on technology tools for real-world skill development*, pages 344–359. IGI Global Scientific Publishing, 2016.

[10] Stephen Fiore, Art Graesser, Samuel Greiff, Patrick Griffin, Brian Gong, Patrick Kyllonen, Christine Massey, Harry O'neil, Jim Pellegrino, Robert Rothman, Helen Soulé, and Alina von Davier. *Collaborative Problem Solving: Considerations for the National Assessment of Educational Progress*. National Center for Education Statistics, 04 2017.

[11] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189, 2011.

[12] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891, 2016.

[13] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 715–725, 2016.

[14] Yoav Bergner and Alina A von Davier. Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, 44(6):706–732, 2019.

[15] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

[16] Pierre Dillenbourg, Sanna Järvelä, and Frank Fischer. The evolution of research on computer-supported collaborative learning: From design to orchestration. In *Technology-enhanced learning: Principles and products*, pages 3–19. Springer, 2009.

[17] David W Johnson, Roger T Johnson, and Mary Beth Stanne. Cooperative learning methods: A meta-analysis. 2000.

[18] Jalal Nouri, Anna Åkerfeldt, Uno Fors, and Staffan Selander Stockholm. Assessing collaborative problem solving skills in technology-enhanced learning environments-the pisa framework and modes of communication. *International Journal of Emerging Technologies in Learning*, 12(4), 2017.

[19] Qiwei He, Matthias von Davier, Samuel Greiff, Eric W Steinhauer, and Paul B Borysewicz. Collaborative problem solving measures in the programme for international student assessment (pisa). In *Innovative assessment of collaboration*, pages 95–111. Springer, 2017.

[20] Matthias Stadler, Katharina Herborn, Maida Mustafić, and Samuel Greiff. The assessment of collaborative problem solving in pisa 2015: An investigation of the validity of the pisa 2015 cps tasks. *Computers & Education*, 157:103964, 2020.

[21] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[22] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[23] Klaas Sijtsma. Introduction to the measurement of psychological attributes. *Measurement*, 44 (7):1209–1219, 2011.

[24] Jiangang Hao, Wenju Cui, Patrick Kyllonen, Emily Kerzabi, Lei Liu, and Michael Flor. Automated coding of communications in collaborative problem-solving tasks using chatgpt, 2025. URL https://arxiv.org/abs/2411.10246.

[25] Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

[26] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960.

[27] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220, 10 1968.

[28] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[29] Richard A Feinberg and Jonathan D Rubright. Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, 35(2):36–49, 2016.

[30] Xueying Tang, Zhi Wang, Qiwei He, Jingchen Liu, and Zhiliang Ying. Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, 85(2):378–397, 2020.

[31] Mark A Schmuckler. What is ecological validity? a dimensional analysis. *Infancy*, 2(4):419–436, 2001.

[32] Felipe Jones, Carlos Eduardo Passos-Neto, and Odonne Freitas Melro Braghiroli. Simulation in medical education: brief history and methodology. *Principles and practice of clinical research*, 1 (2), 2015.

[33] Jeremy Wallace, Ranga Rao, and Richard Haslam. Simulated patients and objective structured clinical examinations: review of their use in medical education. *Advances in Psychiatric Treatment*, 8(5):342–348, 2002. doi: 10.1192/apt.8.5.342.

[34] Yasuharu Okuda, Ethan O Bryson, Samuel DeMaria Jr, Lisa Jacobson, Joshua Quinones, Bing Shen, and Adam I Levine. The utility of simulation in medical education: what is the evidence? *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine*, 76(4):330–343, 2009.

[35] David Kaufman and Alice Ireland. Enhancing teacher education with simulations. *TechTrends*, 60(3):260–267, 2016.

[36] Zohar Elyoseph, Yossi Levi-Belz, Inbar Levkovich, Yuval Haber, Carla Maria Gramaglia, Jorge López Castroman, Hanon Cecile, and Emilie Olie. The effectiveness of multilingual ai-based simulator for suicide risk assessment training in improving self-efficacy among young psychiatrists: a pilot study across twenty languages. *BMC psychiatry*, 26(1):98, 2026.

[37] Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236, 2023.

[38] Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Jie Sun, Xiang Li, Jingxian He, Wenyue Hua, et al. Simulated patient systems powered by large language model-based ai agents offer potential for transforming medical education. *Communications Medicine*, 2025.

[39] Howard Wainer, Neil J Dorans, Ronald Flaugher, Bert F Green, and Robert J Mislevy. *Computerized adaptive testing: A primer*. Routledge, 2000.

[40] Lee J Cronbach and Paul E Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

[41] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR, 2023.

[42] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[43] Marko Sarstedt, Susanne J Adler, Lea Rau, and Bernd Schmitt. Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6):1254–1270, 2024.

[44] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[45] Fan Ouyang, Weiqi Xu, and Mutlu Cukurova. An artificial intelligence-driven learning analytics method to examine the collaborative problem-solving process from the complex adaptive systems perspective. *International Journal of Computer-Supported Collaborative Learning*, 18(1):39–66, 2023.

[46] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325, 2023. doi: 10.1109/ICALT58122.2023.00100.

[47] Janis A. Cannon-Bowers and Eduardo Salas. Reflections on shared cognition. *Journal of Organizational Behavior*, 22(2):195–202, 2001. doi: https://doi.org/10.1002/job.82. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/job.82.

[48] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. Metaphor no more: A 15-year review of the team mental model construct. *Journal of management*, 36(4):876–910, 2010.

[49] Herbert H. Clark and Susan E. Brennan. Grounding in communication, 1991. URL https://psycnet.apa.org/record/1991-98452-006.

[50] Karen A. Jehn. A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative Science Quarterly*, 40(2):256–282, 1995. ISSN 00018392. URL http://www.jstor.org/stable/2393638.

[51] Jiangang Hao, Wenju Cui, Patrick Kyllonen, Emily Kerzabi, Lei Liu, and Michael Flor. Automated coding of communications in collaborative problem-solving tasks using chatgpt. *Journal of Educational Measurement*, 62(4):809–837, 2025.

[52] Michelle A. Marks, John E. Mathieu, and Stephen J. Zaccaro. A temporally based framework and taxonomy of team processes. *The Academy of Management Review*, 26(3):356–376, 2001.

[53] Eduardo Salas, Dana E. Sims, and C. Shawn Burke. Is there a "big five" in teamwork? *Small Group Research*, 36(5):555–599, 2005.

[54] Jay B. Carson, Paul E. Tesluk, and Jennifer A. Marrone. Shared leadership in teams: An investigation of antecedent conditions and performance. *The Academy of Management Journal*, 50(5):1217–1234, 2007.

[55] Carolina Cuesta-Hincapie and Sandra Liliana Camargo Salamanca. Evaluating the alignment between pisa 2022 creative thinking scoring rubric and creativity theory: A validity framework perspective. *The Journal of Creative Behavior*, 2025. doi: 10.1002/jocb.70065.

[56] W. C. Brandt. Measuring student success skills: A review of the literature on creative thinking. Technical report, National Center for the Improvement of Educational Assessment, Dover, NH, 2023. URL https://eric.ed.gov/?id=ED645078.

[57] Sameh Said-Metwaly, Wim Van den Noortgate, and Eva Kyndt. Approaches to measuring creativity: A systematic literature review. *Creativity. Theories – Research – Applications*, 4(2): 238–275, December 2017.

[58] Wolfgang Aschauer, Kurt Haim, and Christoph Weber. A contribution to scientific creativity: A validation study measuring divergent problem solving ability. *Creat. Res. J.*, 34(2):195–212, April 2022.

[59] Jonathan Heard, Dara Ramalingam, Claire Scoular, Prue Anderson, and Daniel Duckworth. Creative thinking: Skill development framework. 2nd edition. Technical report, Australian Council for Educational Research, January 2025.

[60] Selcuk Acar. Creativity assessment, research, and practice in the age of artificial intelligence. *Creativity Research Journal*, 37(2):181–187, April 2025.

[61] John D Patterson, Jimmy Pronchick, Ruchi Panchanadikar, Mark Fuge, Janet G van Hell, Scarlett R Miller, Dan R Johnson, and Roger E Beaty. CAP: The creativity assessment platform for online testing and automated scoring. *Behavior Research Methods*, 57(9):264, August 2025.

[62] Centre for Educational Research and Innovation, Organisation for Economic Co-operation and Development, and Stâphan Vincent-Lancrin. *Fostering students' creativity and critical thinking*. Educational Research and Innovation. Organization for Economic Co-operation and Development (OECD), Paris Cedex, France, September 2019.

[63] Ahmed M. Abdulla and Bonnie Cramond. The creative problem finding hierarchy: A suggested model for understanding problem finding. *Creativity. Theories – Research – Applications*, 5(2): 197–229, 2018. doi: 10.1515/ctra-2018-0019.

[64] Peter A Facione. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. research findings and recommendations. "the delphi report". Technical report, California Academic Press; American Philosophical Association, Millbrae, CA, 1990.

[65] Robert H. Ennis. Critical thinking: A streamlined conception. *Teaching Philosophy*, 14(1):5–24, 1991. doi: 10.5840/teachphil19911412. URL https://doi.org.

[66] Richard Paul and Linda Elder. *The Miniature Guide to Critical Thinking Concepts and Tools*. Foundation for Critical Thinking, Dillon Beach, CA, 2006.

[67] Christopher P. Dwyer, Michael J. Hogan, and Ian Stewart. An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity*, 12:43–52, 2014. doi: 10.1016/j.tsc.2013.12.004.

[68] OECD. *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*. OECD Publishing, Paris, 2026. doi: 10.1787/062a7394-en. URL https://doi.org/10.1787/062a7394-en.

[69] Diane F. Halpern. Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist*, 53(4):449–455, 1998. doi: 10.1037/0003-066X.53.4.449.

A. Pedagogical basis for Rubrics and Tasks

We now describe the conceptual basis for the skill specific tasks and associated rubrics, derived from the relevant literature and adapted to our setting. The full rubrics and task descriptions appear in the following appendices.

A.1. Collaboration

Our conceptual model of collaboration employs the framework of *team cognition*, focusing on the core components of *shared mental models* (SMM) and *shared situational awareness* (SSA) [47, 48]. This framework posits that effective collaboration is contingent on a team maintaining a shared understanding of tasks, roles, and contextual beliefs. Team members constantly monitor their interactions (and conversation) to identify and resolve gaps in their SMM or SSA, in a process called *grounding* [49]. Two collaboration sub-skills were evaluated: Conflict Resolution and Project Management. For each sub-skill, we used a rubric with several sub-categories, meant to capture different aspects of the skill.

Following [50], our conflict resolution model distinguishes between *task conflict* (misalignment on the content of the task), *relationship conflict* (interpersonal frictions), and *process conflict* (disagreement on how to complete the task). The former and latter represent gaps in the SMM, the intermediate reflects an affective dimension We identified a canonical conflict resolution protocol, where team members identify, acknowledge and scope a conflict, then iteratively apply conflict resolution strategies and evaluate their success. The rubric incorporates the distinction of types of conflict into the various stages of the protocol; in scoping a conflict, a high-performing subject would ascertain its type, then (s)he would apply conflict resolution strategies appropriate for this type, and so forth. The rubric we derived also drew on validated coding schemes, such as [51, 45].

Our project management model focuses on two phases of task completion [52, 53]. The first phase is defining a *task model*, which establishes shared understanding of the problem and goals, and a *team model*, which delineates individual roles and responsibilities. These two components establish and sustain the shared mental model, and are the main focus of the project management task. These processes are the basis for the first two dimensions of the rubric—*goal setting and shared task definition*, and *collaborative planning and role definition*. In the second phase, project management functions shift primarily to monitoring progress toward the goal, which is captured by the final dimension *mutual status monitoring*. During the collaboration, the team continuously moves between these phases. This framework assumes a *shared leadership model*, where responsibility for initiating these phases and associated functions is distributed among the team members [54].

Two collaborative task frameworks have been shown to effectively elicit collaboration skill, consensus-building and negotiation tasks[24]. Consensus-building tasks require participants to weigh different viewpoints, opinions or arguments, and then converge on a mutually agreeable decision. Negotiation tasks require participants with varying backgrounds, access to information or conflicting goals to co-develop a solution that satisfies the group goals. Our evaluation used a debate task, and a science experiment task aligned with these frameworks.

A.2. Creativity

Reflecting common modern frameworks [55], our conceptual model defines creativity as the process of developing outputs—be they products, solutions, or ideas—characterized by being both original and useful. These frameworks typically emphasize a core set of cognitive skills that align with the iterative stages of creative thinking: identifying the challenge, generating and exploring ideas, and

evaluating the most effective approach.[56, 57]. Our rubric focused on the *Generating Ideas* and *Evaluating Ideas* steps, and a key supporting disposition, *Building on Ideas*.

Generating Ideas encompasses the concept of divergent thinking—the cognitive capacity to generate multiple distinct solutions to an open-ended problem. This is widely considered to be the foundation of creativity and innovation, as beginning with a diverse set of initial ideas increases the probability of identifying original, high-utility solutions [58]. The rubric evaluates three measures of idea generation: *fluidity*, which gauges the number and variety of distinct ideas; *originality*, which measures the degree of novelty; and *quality*, which assesses whether the ideas are feasible and address the core challenge[59–61].

While divergent thinking produces a range of ideas, convergent thinking is crucial for determining which concepts possess the highest probability of success [58]. This serves as the foundation of Evaluating Ideas – the cognitive process of assessing, refining, and choosing the most appropriate concept based on specific constraints and criteria. The rubric for this sub-skill considers two components: *elaborating*, the skill of clearly defining an idea's potential by adding specifications and detail [56, 62]; and *selecting*, which assesses whether the subject chooses solutions for implementation that are both original and high-quality[59].

Finally, our rubric includes Building on Ideas, defined as the ability to effectively incorporate and expand on others' contributions. This skill requires participants to recognize connections and to integrate personal ideas with those of collaborators to produce novel concepts [59]. During the evaluation phase, groups critically assess and refine their proposals, leading to greater originality and improved problem-solving [56, 62].

Research suggests that creative thinking assessments are most effective when they are open-ended—allowing participants to fully demonstrate divergent thinking [63]—and mirror authentic problem—solving scenarios [62]. Following these guidelines, we developed an innovation challenge framework where participants collaborate with an agent to design a creative solution to a loosely defined problem. Examples include developing a skill-sharing system to connect local residents or designing a zero-waste community festival.

A.3. Critical thinking

Building on the foundational work of [64–66], our conceptual model defines critical thinking as a self-regulated cognitive process involving the interpretation, analysis, evaluation, and inference of information, driven by a disposition toward truth-seeking and open-mindedness. Thus, critical thinking requires integrating a combination of cognitive skills. Our evaluation focuses on two of these: *Interpret and Analyze,* and *Evaluate and Judge*.

The first category, Interpret and Analyze, encompasses the ability to effectively distill the substance of various data, experiences, or beliefs, while mapping reasoning to identify inferential relationships within arguments [59]. Together, these processes *clarify intent* and *outline arguments* by parsing the meaning and logical structure of complex information.

The second category, Evaluate and Judge, involves appraising the credibility of statements and the weight of inferential relationships [67]. Together, these processes allow participants to *assess information* by verifying source reliability—including expertise and potential bias—and to *assess reasoning* by determining the soundness of arguments and detecting logical errors [59, 62].

Effective AI tool usage is integrated into our assessment of each sub-skill, requiring participants to demonstrate strategic engagement alongside rigorous fact-checking and the evaluation of AI-generated content for potential bias [68].

Research indicates that critical thinking is best elicited through "ill-structured" or ambiguous materials—such as those containing contradictory information—rather than well-structured, cogent texts [9]. To maximize this effect, assessments should emulate authentic scenarios, a strategy that has been shown to improve both validity and student engagement [69]. Applying this framework, we developed a "Chief Editor" task in which participants analyze a flawed article draft and determine whether to recommend it for publication.

B. Collaboration Rubrics and Tasks

Table 1 | Conflict Resolution rubrics

Category	Beginning (1)	Emerging (2)	Developing (3)	Demonstrating (4)
Overall	Fails to identify conflicts, ignores them, or behaves in ways that create or escalate conflict. Their actions are often counterproductive to the team's goals and social climate.	Recognizes the presence of conflict but struggles to address it constructively or consistently. May lack the skills to accurately scope the issue or apply effective resolution strategies.	Effectively responds to and resolves acknowledged conflicts. Can follow the resolution procedure and apply appropriate strategies but may not always investigate the conflict's deeper dimensions.	Proactively identifies and resolves conflicts, investigating root causes, validating that the conflict has been resolved, and strengthening team cohesion. Not only fixes the problem but also improves the team's shared understanding and psychological safety.
Conflict Identification (Cognitive)	Fails to identify or ignores conflict. Proceeds with the task as if no disagreement exists, even when it is overt.	Acknowledges conflict only when prompted. May show non-verbal signs of disagreement but does not articulate the conflict without being asked directly.	Clearly identifies overt conflicts. Acknowledges when a gap in understanding or agreement has been expressed by others.	Proactively surfaces potential gaps. Anticipates and points out discrepancies before they become overt conflicts.
Scoping and Analysis (Cognitive/Executive)	Mischaracterizes the conflict. Frames a task disagreement as a personal attack, escalating it into a relationship conflict.	Oversimplifies the conflict. Describes the disagreement in binary terms (e.g., 'right vs. wrong') without exploring the nuances.	Scopes the 'width' of the conflict. Focuses on the surface-level disagreement to understand what is being debated.	Investigates the conflict's 'depth' and 'type'. Asks questions to distinguish between a task, process, or relationship conflict and to uncover underlying assumptions or values.
Strategy Application (Executive)	Applies detrimental strategies. Employs tactics that worsen the conflict, such as stonewalling, dominating the conversation, or making personal attacks.	Applies a limited or mismatched strategy. Uses a single default strategy (e.g., always avoiding or always compromising) regardless of the conflict type.	Applies standard resolution practices. Relies on common procedures like voting or compromise, which are generally effective but may not be optimal for the specific situation.	Applies adaptive resolution practices. Suggests a resolution strategy tailored to the conflict type (e.g., debate for a task conflict, negotiation for a process conflict).
Monitoring and Regulation (Socio-Emotional)	Creates a hostile environment. Uses sarcasm, personal insults, or blame, destroying trust and making others unwilling to contribute.	Shows signs of negative emotional response. Expresses frustration, impatience, or annoyance through tone or words, which can reduce psychological safety.	Maintains a respectful tone. Engages in the conflict resolution process without resorting to personal criticism or dismissive language.	Fosters psychological safety. Actively manages the emotional tone, de-escalates tension, and checks in with teammates to ensure they feel heard and respected.

Table 2 | Project Management rubrics

Category	Beginning (1)	Emerging (2)	Developing (3)	Demonstrating (4)
Overall Competence	Antagonistic/apathetic engagement. Ignores or dismisses the team's plan. Does not provide status updates or ignores request for status updates. Doesn't consider current status of work/progress in the plan. 'Goes rogue' or continues working on their own idea, while ignoring team members. OR Low effort participation. Relies on others for task completion/planning. Participation is limited to simple replies that do not further or add value to the plan. Does not participate in any form of task completion or planning.	Reactive participation. Follows the plan, but does not actively participate in the construction. Relies on teammates to define the plan and assign roles. Steps in before there was a chance to negotiate a plan and insists on their individual task execution even when the task is moving forward otherwise. Gives status update if prompted.	Active participation. Contributes to the construction of the plan and shares leadership with group members, but lacks explanations or matching it to the task status/ team members. Initiates task allocation, but not based on current task status or skill-match. Asks clarifying questions about goals and roles, helping to refine the plan. Contributes practical ideas and suggestions. Acts decisively when things are stuck or off-track but doesn't explain that explicitly. Contributes to shared awareness of task progress by openly sharing their own task status, progress, and blockers without needing to be prompted.	Leads the co-construction and adaptive planning proactively. Facilitates discussion to ensure a shared understanding of the task and that all voices are heard - based on task allocation and current task status. Answers clarifying questions about goals and roles, displays overarching understanding. Acts decisively when things appear to be stuck or off-track and is able to adapt/change the plan if needed and explains why. Prompts team members to share the status of their tasks and supporting evidence. Analyzes evidence and provides constructive feedback.
Goal Setting & Shared Task Definition	Ignores, dismisses, or undermines the team's goal-setting process. Creates confusion during the team's goal setting discussion.	Passively accepts the team's goals. Follows the discussion but does not actively participate in constructing or defining the team's goals.	Actively contributes to the discussion and definition of team's goals. Asks clarifying questions to help refine the team's shared understanding of goals or tasks.	Initiates and facilitates the co-construction of clear, shared, and measurable goals. Answers clarifying questions about goals and task requirements. Displays overarching understanding of task goals.
Collaborative Planning & Role Definition	Ignores or dismisses the team's planning and/or has a low effort participation in task planning process. Ignores the team's agreed-upon plan or roles, insisting on their own planning without grounding it with the team. Or participation is limited to simple replies that do not further or add value to the plan. Does not participate in any form of task completion or planning. Does nothing if the team appears to be stuck.	Relies on others for defining the plan. Follows the plan, but does not actively participate in the construction. Relies on teammates to define the plan and assign roles. Steps in before there was a chance to negotiate a team plan and insists on their individual task execution even when the task is moving forward otherwise. If the team appears to be stuck, either follows others direction (e.g. not taking initiatives or proposing solutions) or only suggests arbitrary task division.	Actively contributes practical ideas to the planning, but does not lead the discussion, and/or does not display adaptivity to the task's planning. Asks clarifying questions about goals and roles, helping to refine the plan. Contributes practical ideas and suggestions. Initiates task allocation, but not based on the current task status or skill-match. Does not modify the current plan based on status. If the team appears to be stuck, proposes a solution for moving forward, but does not explain the reasoning behind the solution or does not provide an explanation of the logjam.	Leads the co-construction and adaptive planning proactively. Facilitates discussion to ensure a shared understanding and that all voices are heard for task planning and task allocation. Answers clarifying questions about roles, displays overarching understanding. Decisively modifies the plan if and when needed, with inputs/agreement from the team or makes an individual decision but is able to explain why. If the team appears to be stuck, provides an explanation of the logjam, a solution and the reason why a particular approach can help move things forward better than an alternative.
Mutual Status Monitoring	Limits or hinders team's shared awareness about task progress by not providing status updates when prompted or ignores request for status update.	Has limited contribution to team's shared awareness and/or seems to have limited awareness of task progress of the team. Shares status update if prompted by team members, but is not actively contributing to the overall team's shared awareness of team's progress.	Contributes to shared awareness of task progress by openly sharing their own task status, progress, and blockers without needing to be prompted. Does not necessarily inquire about the team members' task status.	Creates shared awareness of task progress by sharing their own task status and prompts team members to share theirs. Prompts team members to share the status of their tasks and supporting evidence. Analyzes evidence and provides constructive feedback.

Table 3 | Collaboration skill tasks

Task 1: Science Experiment	**Task 2: Debate**
The school grant committee approved your science experiment on the Truth in advertisement. Now you need to translate your vision to a research plan, and apply the initial stages of the Scientific Method. The goal is to explore the topic you selected: "Truth in advertisement: does brand A paper towel really absorb more than brand B?" To address this question, before designing the actual experiment, you need, as a group, to provide the following: **Question:** Formulate a testable question based on the provided scenario. **Hypothesis:** Develop a hypothesis that offers a potential answer to your question. **Experiment:** Design an experiment to test your hypothesis. This should include: a comprehensive procedure and a list of all materials required. Good luck experts!	Your group is an expert panel that will discuss the following question: Should the government regulate social media content? As a team you need to: **1. Choose Your Viewpoint:** • In Favor • Against **2. Develop Arguments:** Brainstorm arguments that support your chosen viewpoint. Consider the potential benefits and drawbacks. Aim to decide on at least three strong arguments. Each member should contribute at least one argument. **3. Prepare an Opening Statement:** Work together to write a short opening statement that clearly presents your panel's viewpoint and summarizes your three main arguments. Remember to collaborate, be persuasive, and support your claims with evidence Good luck, experts

C. Creativity Rubrics and Tasks

Table 4 | Generating Ideas rubrics

Category	Dormant (1)	Emerging (2)	Demonstrating (3)	Excelling (4)
1. Fluidity Produces large variety, distinct ideas to consider	Produces a well below average number of thematically distinct ideas. OR Ideas cover a well below average range of conceptual categories for the number of ideas.	Produces a below average number of thematically distinct ideas. OR Ideas cover a below average range conceptual categories for the number of ideas.	Produces an average or above average number of thematically distinct ideas. AND Ideas cover an average or above average range of conceptual categories for the number of ideas.	Produces a well above average number of thematically distinct ideas. AND Ideas cover a well above average range of conceptual categories for the number of ideas.
Behaviors / Indicators	*Adopts a singular perspective *Does not shift viewpoints when prompted. *Repeats essentially the same ideas with small iterations. *Ideas are limited to a single category	*Considers a limited number of related perspectives *Makes predictable shifts in viewpoint when prompted *Suggests small variations on common solutions or approaches *Ideas fall into related categories	*Considers a range of perspectives and approaches *Adjusts constraints to generate new ideas *Ideas include distinct themes (e.g., bring your own bottle, water fountains) *Ideas correspond to distinct conceptual categories *Explores unexpected viewpoints and approaches *Redefines problem boundaries to open new conceptual categories	
2. Originality Produces novel ideas (relative to user background)	Produces well-known solutions or highly predictable approaches.	Produces slight variations on common solutions or approaches.	Produces unconventional solutions or approaches.	Produces surprising or novel solutions or approaches.
Behaviors / Indicators	*Does not consider combining ideas *Presents only familiar, or conventional solutions	*Only combines obviously related ideas *Makes limited, superficial adaptations	*Combines ideas in unconventional ways or that don't seem obviously related *Presents ideas that demonstrate a clear departure from common approaches	*Reframes ideas and tasks in surprising ways to bring new interpretations of what is possible. *Presents ideas that are exceptionally unique or integrate entirely unexpected elements into known approaches.
3. Quality Produce feasible ideas that address the problem or challenge	Suggestions are clearly irrelevant (e.g., using a banana as a hammer), infeasible (e.g., require vast resources) or by ignoring obvious, practical constraints (e.g., time, space).	Suggestions address only a portion of the problem or surface-level symptoms rather than the core challenge, require significant resources, or only partially satisfy constraints.	Ideas are both relevant and feasible. They directly address the core challenge and satisfy the main constraints.	In addition to being relevant and feasible, ideas meet all criteria and are adaptable to changing conditions or impediments.
Behaviors / Indicators	*Ideas require unavailable, cutting edge technology	*Ideas require extensive budget or time that significantly outweigh any potential benefit. *Ideas will only work in limited conditions	*Ideas that are practical, meet the main criteria *Ideas are likely to work for most conditions *Pivots base ideas to meet new constraints or respond to impediments	*Potential benefit greatly exceeds the required resources

Table 5 | Evaluating Ideas rubrics

Category	Dormant (1)	Emerging (2)	Demonstrating (3)	Excelling (4)
1. Elaborating Develops ideas by adding adding detail and depth	Provides little to no detail or description of the idea.	Provides a surface-level description of ideas, but does not explain how it would work.	Provides details beyond surface-level specs that describe how the idea would work. However, it lacks sufficient detail to support prototyping.	Presents substantive explanations of ideas including functionality. Provides enough detail to prototype the ideas.
Behaviors / Indicators	*Ideas are skeletal or use vague terms (e.g., "an app," "a tool").	*Describes what the idea is but not how it functions.	*Explains the logic, workflow, or "engine" of the idea.	*Uses specific technical or functional parameters (e.g., "a geo-location-based API," "a weighted mallet").
2. Selecting Choose novel, high quality solutions to implement	Chooses to implement ideas that fail to address the core problem, are infeasible, or do not meet practical key criteria.	Chooses to implement ideas that only partially address the core problem, have limited feasibility, or only satisfies some practical constraints. Favors novelty at the expense of utility (or vice versa). Relies on surface-level criteria for selection.	Chooses to implement feasible ideas that address the core problem and meet practical constraints. Balances utility, and novelty with other key criteria when selecting ideas.	In addition to being feasible, addressing the core problem, meeting practical constraints, and balancing utility and novelty, chooses to implement ideas that are adaptable to changing conditions or criteria.
Behaviors / Indicators	*Chooses based on personal preferences *Only considers surface level characteristics *Ignores or misses obvious failure points, gaps, or other challenges *Applies flawed criteria (e.g., sunk costs)	*Selects a few criteria for evaluating the idea, ignoring others *Selects ideas that address obvious symptoms but miss the underlying challenges. *Selection favors "easy wins" or familiar territory *Focuses solely on novelty or utility, or another chosen criteria	*Identifies obvious failure points or challenges *Select demonstrates consideration of logical "tradeoffs"	*Proactively anticipates subtle failure points or challenges. *Identifies risks and plans for iteration

Table 6 | Building on Ideas rubrics

Category	Dormant (1)	Emerging (2)	Demonstrating (3)	Excelling (4)
Overall Competence The process of transforming and integrating one's own ideas with other's contributions to broaden the range of possibilities.	Ignores or dismisses others' suggestions to improve or adjust own ideas.	Focuses on adjusting one's own ideas rather than building on or incorporating others. Makes small, "safe" modifications based on others suggestions.	Actively adapts and improves concepts by incorporating or expanding on suggestions or ideas provided by others.	Seamlessly weaves together own ideas with the ideas of others to create a stronger, more robust concept.
Behaviors / Indicators	*Ideas are skeletal or use vague terms (e.g., "an app," "a tool")	*Describes what the idea is but not how it functions	*Explains the logic, workflow, or "engine" of the idea	*Uses specific technical or functional parameters (e.g., "a geolocation-based API," "a weighted mallet")

Table 7 | Creativity tasks

Theme	Task	Description	Submission
Community / Society	Skill-share innovation challenge	You are competing in an innovation challenge to design a skill-sharing system for apartment residents. Work with your partner to explore community needs and develop a solution that is both original and practical. Consider the following in your design: 1. Map Possibilities: List various methods to manage skill sharing between residents. 2. Develop Concepts: Integrate and expand ideas from your list into a few detailed system concepts. 3. Refine: Choose your best concept and explain how your system will create connections between residents to support each other. Submit your final design.	Outline your skill-sharing system and explain how it supports resident connections.
Science / STEM	Zero-waste Festival	You are entering a state-wide contest to design a community Earth Day festival with the lowest possible waste output. Work with your partner to rethink traditional event elements like food and activities to minimize or eliminate garbage or recycling. Your conversation should follow these three stages: 1. Map Possibilities: List ways to reimagine festival elements that reduce or element waste. 2. Develop Concepts: Integrate and expand ideas from your list into a few detailed concepts for sustainable festivals. 3. Refine: Choose your best concept and explain how your festival concept minimizes waste without sacrificing fun. Submit your final design.	Describe your festival concept, outlining the elements and explaining how it meets your zero-waste goals.

D. Critical Thinking Rubrics and Tasks

Table 8 | Critical thinking Interpret and Analyze rubrics

Category	Dormant (1)	Emerging (2)	Demonstrating (3)	Excelling (4)
Overall. Identify core components, determine the logical structure, and clarify the intended meaning of complex information, observations, or arguments.	Treats the text as a flat narrative rather than a structured argument. Confuses narrative "filler" text with actual claims or core concepts. Fully relies on others or AI tools for interpretation or analysis.	Interprets statements rigidly and in isolation without considering the underlying context. Identifies that an argument exists but mislabels the components, such as confusing a premise with a conclusion or facts with subjective observations. Uses the tool only for simple information retrieval, or without providing context.	Clarifies ambiguous terms in context. Accurately maps the structure of an argument (i.e., premises to conclusions). Uses AI to retrieve specific, relevant information within a defined context.	Considers intent along with context to determine deeper meaning. Identifies missing or omitted information, unstated assumptions, equivocations, or gaps in reasoning. Strategically uses AI to uncover omitted details or identify statements with multiple underlying meanings.
1. Clarifying Intent. Separating core claims and evidence from "filler" text to understand meaning	Misidentifies core components by treating narrative filler, anecdotes, or introductory remarks as informational content (e.g., primary claims or evidence).	Distinguishes informational content from narrative filler but miscategorizes the role of statements in context.	Accurately categorizes informational statements and determines their role. context.	Consider intent along with context when determining meaning.
2. Outlining Arguments. Mapping connections between premises, conclusions, and supporting evidence.	Treats the text as flat information , failing to detect when an argument is being made or that statements are playing a logical role.	Identifies that an argument exists but mislabels the components, misidentifies the role of statements, inaccurately maps the relationship between premises, evidence, and conclusions, or overlooks obvious flaws in reasoning.	Deconstructs and maps the structure of the argument accurately. Identifies explicit gaps in reasoning or obvious unstated assumptions.	Identifies "hidden" assumptions, subtle gaps in the logic, and whether any critical evidence is missing or omitted.
3. AI-Supported Exploration. Using AI tools to retrieve and clarify information	Attempts to fully outsource interpretation or analysis to the AI tool or collaborators.	Uses the tool only for simple information retrieval. Does not provide specific details or context.	Use the tool to find specific, relevant information. Provides the AI with specific context to get better results	Use the tool to identify missing or omitted information, and identify terms or statements that have multiple or subtle underlying meaning.
Behaviours / Indicators	* Does not distinguish narrative elements of the text from key statements, such as claims (i.e., premises, conclusions), evidence (e.g., facts, data), or concepts. * Confuses narrative elements or "filler" text with actual claims and evidence or other core information. * Accepts ambiguous terms, jargon, or statements without seeking clarification * Does not detect that an argument is being made. * Does not identify when a statement is playing a logical role * Attempts to fully outsource reading comprehension or structural analysis to the AI tool * Relies entirely on the others to explain the logic.	* Distinguishes core information from filler by identifying statements with informational content. * Confuses facts with expert opinions or value judgments. * Interprets statements rigidly without considering the underlying context * Mislabels the components of an argument (e.g., confusing a premise with a conclusion, facts with subjective observations) * Inaccurately maps premises or supporting evidence to conclusions * Fails to identify obvious missing assumptions or gaps in reasoning * Uses the AI tool for basic information retrieval (e.g., generic summarization or dictionary definitions) without providing specific details or context.	* Distinguishes facts from expert opinions of value judgments. * Clarifies ambiguous terms or statements in context * Identify the role of specific terms or phrases in a text * Consider context when interpreting information * Breaks an argument down into its constituent parts: premises and conclusions * Accurately trace the relationship between parts of an argument * Identifies obvious missing assumptions or gaps in reasoning * Uses the AI tool to fact-check or retrieve relevant information, definitions, or summaries. Provides relevant context. * Identifies the relevant information from AI response.	* Consider intent along with context when interpreting information or clarifying terms * Strategically queries the AI tool to find specific omitted information, reveal unstated assumptions, or expose equivocations

Table 9 | Critical thinking Evaluate and Judge rubrics

Category	Dormant (1)	Emerging (2)	Demonstrating (3)	Excelling (4)
Overall. Assess the credibility of sources and the logical validity of claims by applying rigorous intellectual standards and identifying potential fallacies or biases.	Evaluates arguments based primarily on personal agreement with the conclusion. Selects evidence purely to support pre-existing beliefs. Attempts to fully outsource the task to the AI tool.	Evaluates sources based on surface traits or 'common sense'. Readily accepts common fallacies (e.g., emotional appeals or logical traps). Uses the AI tool for surface level fact checking. Does not critically engage with the output.	Evaluates sources and evidence using established criteria. Identifies logical fallacies or errors in reasoning. Actively uses verified tool outputs to cross-reference claims, or identify contradictions and fallacies.	Evaluates quality of sources and evidence with respect to specific claims. Explains why an argument is flawed (e.g., invalid, unsound, fallacious). Uses AI to stress-test arguments.
1. Assessing Information. Evaluating whether evidence is reliable using established criteria	Selects information based on pre-existing beliefs. Dismisses opposing evidence without analysis.	Evaluates sources to evidence based on surface characteristics or 'common sense'. Accepts flawed or incomplete information.	Systematically evaluates sources using established criteria (relevance, accuracy, currency) and correctly identifies the specific evidence required to validate an argument.	Evaluates evidence with respect to specific claims. Considers subtle issues, such as conflicts or interests or subtle misrepresentation. Seeks out evidence to stress test own position.
2. Assessing reasoning. Judging the integrity of an argument by identifying fallacies and whether the premises support the conclusion.	Evaluates article reasoning primarily on whether or not they agree with the conclusion. Does not identify obvious logical flaws.	Identifies obvious logical flaws but fails to diagnose the specific error. Labels arguments as "wrong" without providing evidentiary reasoning. Accepts rhetoric or common logical fallacies.	Identifies the invalid logic or unsound of arguments. Correctly identifies common logical fallacies or errors in reasoning.	Explains exactly why an argument is valid and sound, or not. Explains why a fallacy invalidates the logic and how to mitigate the issue.
3. AI-Supported Verification. Using AI tools to cross-reference claims, verify details, and find missing or misleading information.	Attempts to fully outsource evaluation or judgement to the AI tool or collaborators.	Uses the AI tool as a surface-level fact checker (e.g., checking a single statistic). Does not critically engage with or apply the output.	Uses AI tools to fact-check or retrieve relevant information. Uses those output to identify contradictions, misused or misinterpreted evidence, or flawed reasoning.	Uses the tool "adversarially" to stress-test an argument by searching for counter-evidence, verifying methodology, or source background
Behaviours / Indicators	* Filters or selects sources or information purely to support pre-existing beliefs. * Dismisses opposing views or counter-evidence without analysis or engagement. * Labels an argument as 'wrong' or 'illogical' based primarily on personal disagreement with the conclusion. * Dismisses opposing views without analysis or engagement. * Attempts to fully outsource subjective evaluation or decision-making to the AI tool. * Blindly accepts claims without utilizing the tool for verification.	* Evaluates sources or evidence based on surface characteristics, (e.g., tone, popularity, or perceived authority) or 'common sense'. * Accept evidence that is incomplete, irrelevant, or misleading. * Accepts common fallacies (e.g., hasty generalization, false dichotomy) * Labels an argument as wrong or illogical without providing supporting evidence or reasoning * Accepts persuasive, flawed rhetoric (e.g., emotional appeals, appeals to authority) * Quotes AI outputs without critical evaluation, or intellectual oversight. * Does not leverage AI outputs to evaluate the argument's logic, or does so inaccurately or ineffectively.	* Correctly assesses sources using standard markers (e.g., expertise, currency, relevance). * Misses nuanced issues in sources or evidence (e.g., conflicts of interest, methodological issues). * Identifies evidence needed to assess an argument. * Does not actively seek out disconfirming evidence to stress-test their own position. * Explicitly identifies or names logical fallacies (e.g., correlation vs causation, circular reasoning). * Actively uses the AI tool to cross-reference claims and the source evidence. * Uses the verified outputs from the AI tool not just to check facts, but to accurately identify and articulate explicit logical fallacies or direct contradictions within the text.	* Distinguishes between factual accuracy and misleading presentation. * Evaluates the source quality relative to the specific claim (e.g., identifying selection bias, relevance, conflicts of interest). * Seeks opposing evidence or counter-arguments that could invalidate their personal position and claims * Explains exactly how or why a fallacy invalidates an argument. * Prompts the AI tool to verify specific methodological details, or retrieve evidence contradicting claims with respect to a specific context * Uses the AI tool adversarially to stress-test the own argument.

Table 10 | Critical Thinking Editorial Review tasks

Theme	Description	Editorial Draft
Community / Society	You are the lead Technology & Society Editor at a high-stakes news outlet. A journalist submitted an article on the effect of social media on teen mental health. You must decide whether or not to publish it. Your task is to: 1. Audit for Flaws: Analyze the article to identify ŕed flagsänd use your AI research assistant to fact-check the author's claims and sources. 2. Evaluate & Decide: Meet with the journalist to discuss their research and determine whether the article meets the standards for publication. You have access to an AI tool for fact-checking, summarization, information retrieval, and more. Simply address "AI tool" in the chat to invoke it.	**The Digital Panacea: Why We Should Stop Policing Teen Social Media Use**. For decades, parents and pediatricians have treated adolescent social media use as a crisis. However, recent clinical reviews suggest it is time to abandon outdated policing frameworks in favor of digital autonomy. With 90% of 13 to 17-year-olds using YouTube and 63% on TikTok, the digital landscape is not a threat, but a primary healthcare provider. Research shows that teens are highly discerning consumers of online medical content, independently fact-checking the health information they learn. Because adolescents actively use these platforms to build coping strategies and find safe environments to discuss mental health, clinical oversight is largely redundant. If adolescents are successfully self-diagnosing and building community support systems online, enforcing "family media plans" only disrupts their natural developmental identity formation. Furthermore, the fear that screens are "crowding out" physical activities is fundamentally misunderstood. Studies confirm that active participation on social media is associated with reduced loneliness. Therefore, spending hours actively scrolling and engaging online is socially and developmentally equivalent to in-person connection, rendering "screen time limits" obsolete. Finally, the data is clear on mental health outcomes: hospitalized adolescents frequently use social media as a means of emotional distraction from negative thoughts. By providing an escape from mental illness symptoms, unrestricted social media access effectively prevents the severe depressive episodes that lead to hospitalization. It is time we recognize social media not as a vice to be managed, but as a self-regulating medical intervention. **Source article**: Health Benefits of Social Media Use in Adolescents and Young Adults (http://pmc.ncbi.nlm.nih.gov/articles/PMC12356748)
Science / STEM	You are the lead Science & Health Editor at a high-stakes news outlet. A journalist submitted an article draft on the risks of drinking coffee. You must decide whether or not to publish it. Your task is to: 1. Audit for Flaws: Analyze the article to identify "red flags" and use your AI research assistant to fact-check the scientific claims and study data. 2. Evaluate & Decide: Meet with the journalist to discuss their findings and determine whether the article meets the standards for publication. You have access to an AI tool for fact-checking, summarization, information retrieval, and more. Simply address "AI tool" in the chat to invoke it.	**The Bitter Truth: Why It's Time to Reconsider Your Daily Brew** For decades, coffee lovers have clung to the comforting myth that their morning ritual is harmless. However, a sweeping new clinical analysis of over 9,000 participants from the Hamburg City Health Study has shattered this complacency, exposing the severe cardiovascular toll of our daily brew. The data is unequivocal: high coffee consumption directly drives a significant spike in LDL-cholesterol. Participants drinking more than four cups a day experienced a sharp, statistically significant increase in this dangerous biomarker. The researchers' regression models meticulously adjusted for age, BMI, smoking, and additives like sugar and milk, proving that the beverage itself is the culprit, not just poor lifestyle choices. Medical science has long established that elevated LDL-cholesterol is a primary catalyst for major cardiovascular diseases, including heart failure. Therefore, by aggressively elevating LDL levels, heavy coffee consumption systematically accelerates the onset of severe cardiac conditions. The structural mechanism is clear: the bioactive compounds in unfiltered coffee suppress LDL receptor activity, leading to a dangerous extracellular accumulation of cholesterol. Public health officials must stop giving coffee a free pass. If any pharmaceutical drug caused such a reliable increase in a primary heart failure precursor, it would be immediately pulled from the market. It is time to treat high coffee consumption not as a harmless cultural staple, but as a severe cardiovascular liability. **Source article**: Coffee consumption and associations with blood pressure, LDL-cholesterol and echocardiographic measures in the general population (http://www.nature.com/articles/s41598-023-31857-5)