The Narrative AI Advantage?
A Field Experiment on Generative AI-Augmented Evaluations of Early-Stage Innovations

Jacqueline N. Lane[1,2*†], Léonard Boussioux[2,3,4†], Charles Ayoubi[1,2], Ying Hao Chen[3], Camila Lin[3], Rebecca Spens[5], Pooja Wagh[5], and Pei-Hsin Wang[3]

[1]Harvard Business School & Digital Data and Design (D^3) Institute at Harvard
[2]Laboratory for Innovation Science at Harvard (LISH)
[3]University of Washington, Michael G. Foster School of Business
[4]University of Washington, Paul G. Allen School of Computer Science & Engineering
[5]MIT Solve

*Corresponding author: jnlane@hbs.edu

† Jacqueline N. Lane and Léonard Boussioux share co-first authorship.

Abstract

The rise of generative artificial intelligence (AI) is transforming creative problem-solving, necessitating new approaches for evaluating innovative solutions. This study explores how human-AI collaboration can enhance early-stage evaluations, examining the interplay between objective criteria, which are quantifiable, and subjective criteria, which rely on personal judgment. We conducted a field experiment with a global social innovation initiative involving 72 experts and 156 community screeners who evaluated 48 solutions for a global health equity challenge. Screeners received assistance from GPT-4, offering recommendations and, in some cases, a rationale. We compared a human-only control group with two AI-assisted treatments: a black box AI and a narrative AI with probabilistic explanations justifying decisions. Our findings show that AI-assisted screeners were 9 percentage points more likely to fail a solution. There was no significant difference between the black box and narrative AI conditions for objective criteria. However, screeners adhered to narrative AI's recommendations 12 percentage points more often than the black box AI's for subjective criteria. These effects were consistent across both experts and non-experts. Further investigating the attitude of screeners using mouse tracking data, we found that deeper engagement with AI's objective failure recommendations led to more overrides, especially in the narrative AI condition, indicating increased scrutiny. This research underscores the importance of developing *AI interaction* expertise in creative evaluation processes that combine human judgment with AI insights. While AI can standardize decision-making for objective criteria, human oversight and critical thinking remain indispensable in subjective assessments, where AI should complement, not replace, human judgment.

Keywords: Creative evaluation, human-AI collaboration, large language models, generative AI, screening, subjectivity, innovation, AI decision-support, field experiment, social impact

1. Introduction

Organizations increasingly integrate artificial intelligence (AI) into their problem-solving activities, reshaping how knowledge workers generate and evaluate new ideas (Wu et al. 2024). Humans now collaborate with AI on various creative tasks, such as drug discovery (Lou and Wu 2021), image generation and art (Zhou and Lee 2024), market research (Brand et al. 2023), customer interactions (Jia et al. 2023), academic writing (Liang et al. 2024a, Liang et al. 2024b), and creative idea generation (Boussioux et al. 2024, Wang et al. 2023b). These capabilities, fueled by the rise of generative AI, a type of AI technology capable of producing new content, such as text, images, audio, or videos, based on patterns learned from existing data, show tremendous opportunities for creative complementarity between AI and humans (Lebovitz et al. 2022, Raisch and Fomina 2023). Indeed, AI-generated text now often surpasses human capabilities in terms of performance (Brynjolfsson et al. 2023, Dell'Acqua et al. 2023) and online content production (Burtch et al. 2024). Moreover, the proliferation of web resources to aid with AI content creation has lowered the barriers to entry into new venture competitions, crowdsourcing, and crowdfunding platforms.[1] However, the rise of AI-assisted idea-generation capabilities presents a new challenge: the need for effective initial screening processes to separate quality ideas from those that fall short of these standards or are misaligned with organizational goals (Bell et al. 2024). This paper investigates the transformative impact of generative AI, particularly large language models (LLMs), on creative idea evaluation through human-AI collaboration. We explore how LLMs, trained on extensive textual data, reshape decision-making in tasks involving objective and subjective assessment criteria. Our study examines how LLM-generated insights, which structure and rationalize complex decisions, influence human judgment in creative evaluation.

Although human-AI collaboration using LLM technologies is gaining significant traction in idea generation (Boussioux et al. 2024, Wang et al. 2023b), its potential in creative evaluation remains largely unexplored. This imbalance in focus overlooks a critical component of the problem-solving process where AI could potentially offer substantial value. For instance, venture capital firms consider, on average, 101 opportunities for each deal they eventually close. A typical deal takes 83 days to close and an average of 118 hours on due diligence during that period (Gompers et al. 2021). Similarly, organizations involved in distant searches often encounter many innovative suggestions that surpass their management capacity. As a result, decision-makers resort to heuristics and shortcuts to alleviate the cognitive burden of making selection decisions (Piezunka and Dahlander 2015). Organizations must often screen out numerous ideas before considering the "best" ideas to thoroughly evaluate (Bell et al. 2024, Hammedi et al. 2011). LLMs

[1] https://venturebeat.com/ai/how-generative-ai-can-democratize-content-creation-across-the-enterprise/

have tremendous potential to augment the initial screening stage of creative evaluation because of the high volume of solutions of considerable quality at the initial submission stages.

Creative evaluation departs from contexts where human-AI decision-making is typically studied, as it often involves considering objective and subjective criteria. In creative evaluations, the best ideas often emerge from evaluation processes that combine the opinions of multiple evaluators, including experts and non-experts (Boudreau et al. 2016, Kim and DellaPosta 2022, Lane et al. 2023, Mollick and Nanda 2016), as opposed to a correct, ground truth response (Agarwal et al. 2023, Balakrishnan et al. 2022). Whereas objective criteria involve quantifiable and measurable facts, subjective tasks are open to interpretation and based on personal opinion and intuition, including gut instincts (Castelo et al. 2019, Huang and Pearce 2015). Furthermore, the inherently subjective nature of creative content often makes its quality difficult to measure or infer even after the idea is implemented (Boudreau et al. 2016, Criscuolo et al. 2017).

As AI-assisted idea generation becomes increasingly cost-effective and scalable (Boussioux et al. 2024, Girotra et al. 2023), the importance of effective idea screening grows as a mechanism to winnow down the set of alternatives that are being considered for limited organizational funding and resources (Bell et al. 2024, Huang et al. 2023). Integrating LLMs into creative evaluation opens up a new frontier in human-AI collaboration, offering unprecedented opportunities alongside unique challenges. AI systems can augment human capabilities by rapidly processing and analyzing vast amounts of data, applying evaluation criteria consistently across numerous ideas (Lou and Wu 2021). One of the significant advantages of LLMs is their ability to generate "narratives" justifying the choice by producing new text resembling natural human language, which can help clarify and rationalize complex decisions. This computational power can complement human evaluation by reducing fatigue-induced inconsistencies and offering a systematic approach to identifying patterns that might elude human observers.

However, human-AI collaboration in creative evaluation has its drawbacks. While LLMs can rapidly process information, they may lack the nuanced understanding of context that humans bring to evaluations, including cultural, historical, or emotional significance (Amabile 1983, Csikszentmihalyi 1999). Creative evaluation is a task that often requires domain expertise (Boudreau et al. 2016), and an LLM's effectiveness is fundamentally limited by the quality and comprehensiveness of its training data, which may not fully capture the complexity of subjective human experiences.

In a human-AI collaborative setting, the degree of reliance on AI introduces several considerations. On the one hand, humans may over-rely on AI-generated evaluations, potentially diminishing the crucial role of human intuition and judgment in the creative evaluation process (Dell'Acqua 2022, Jacobs et al. 2021). On the other hand, humans, especially domain experts, may reject AI recommendations because of their extensive knowledge of field-specific nuances and higher confidence in their own judgment (Allen and Choudhury 2022, Burton et al. 2020, Krakowsky et al. 2023). Experts may be particularly inclined to discard AI suggestions when disagreeing with the provided rationale, as their expertise allows them to identify flaws in the AI's logic (Bayer et al. 2022, Chen et al. 2023). Alternatively, their expertise might enable them to interpret and utilize AI recommendations more effectively than those without domain-specific knowledge. The circumstances under which they accept or reject AI suggestions and the factors influencing these decisions warrant further investigation.

Furthermore, the nature of LLM training data significantly influences the narrative it provides. For objective criteria, LLMs can often provide verifiable responses. In contrast, for subjective criteria, LLM outputs may reflect a diversity of perspectives present in the training data, potentially affecting how users interact with and interpret these AI recommendations. That being said, the AI's ability to generate confident-sounding narratives, even for subjective matters, could influence both experts' and novices' decision-making processes in ways that are not yet fully understood.

Building on the emerging literature on human-AI collaboration in creative evaluation, our study investigates two critical dimensions affecting the decision process (Figure 1). First, we explore the impact of the delivery mode of AI recommendations, with or without LLM-generated narratives that explain or justify the AI's decision. Second, we examine how the type of task considered affects how evaluators utilize these recommendations, distinguishing between objective (fact-based) and subjective (opinion-based) criteria. These two dimensions lead to three interconnected research questions (RQ):

RQ 1: How do AI-generated narratives influence evaluators' decision-making?

RQ 2: How do objective and subjective decision criteria moderate this influence?

RQ 3: How does domain expertise shape the integration of AI recommendations and narratives in decision-making?

------------ Figure 1 about here ------------

To investigate these research questions, we partner with MIT Solve, an entrepreneurship platform focused on social impact that launches annual global challenges with up to $1 million in funding. We designed a field experiment with the program's directors to investigate how 72 experts and 156 community screeners use generative AI recommendations to support their evaluation decisions on 48 real-world solutions submitted to the 2024 Global Health Equity Challenge. We focused on the first evaluation stage, in which evaluators used a standardized rubric containing objective and subjective criteria to decide whether a solution met a minimum quality threshold to advance to the next evaluation stage. We measured domain expertise based on whether screeners were affiliated with MIT Solve, either as current employees, financial sponsors, or reviewers. We prompted the GPT-4 LLM to recommend whether to pass or fail each idea and provide criteria for failure.

To facilitate the experiment, we developed a custom web application that hosts both the application materials and AI recommendations. Our experimental design included three conditions: a human-only control condition, with no AI assistance; treatment 1: black box AI (BBAI), AI recommendations without rationale; treatment 2: Narrative AI (NAI), AI recommendations with rationale. Importantly, these LLM-generated rationales differ from traditional explainable AI (Barredo Arrieta et al., 2020). Rather than elucidating the model's internal workings, they provide human-like explanations based on patterns in the training data. This distinction allowed us to explore how persuasive, yet potentially unfounded, AI-generated justifications influence human decision-making in creative evaluation. These rationales may allow decision-makers to compare their intuitions, beliefs, or heuristics with the AI narratives, potentially influencing their use of LLM recommendations (Das and Rad 2020). Screeners were randomly assigned to two of the treatment conditions with random picks from the solutions, allowing us to exogenously treat them to different levels of AI support in their decision-making process and derive causal estimates of treatment effects.

Our findings reveal that while screeners consistently utilize AI recommendations to support their decision-making, the extent of this reliance varies depending on the type of recommendation and whether rationale is provided. Notably, screeners were 9 percentage points more likely to fail a solution under the treatment conditions (both BBAI and NAI) than the control condition. When examining the impact of AI recommendation types, we observed that the effects of the BBAI and NAI treatment conditions did not differ when AI recommended passing a solution or failing it based on objective criteria. However, a key distinction emerged with subjective criteria: screeners were more influenced in the NAI treatment condition where the LLM provided plausible narratives for its assessments. In these cases, the NAI treatment led to more significant treatment effects compared to the BBAI condition. Importantly, these effects were consistent across experts and non-experts, highlighting the broad applicability of the treatment conditions.

We analyzed mouse tracking data to understand these behaviors further, which provided insights into how screeners' engagement levels differed depending on AI recommendations. On average, when screeners engaged more deeply with AI recommendations—evidenced by increased mouse clicks—to fail based on objective criteria, they were more likely to override the AI's decision, particularly in the NAI treatment condition. This offers early evidence that deeper engagement may reflect skepticism or a desire to gather additional information to justify a decision contrary to the AI's objective failure assessment.

These findings suggest that LLM-generated narratives, while persuasive, can sometimes lead screeners to accept AI recommendations without critically evaluating the underlying assessments. This deference occurs even though LLMs do not genuinely comprehend the distinction between fact and opinion but rather reflect patterns learned from their training data. The LLM's ability to produce confident and coherent rationale, even on subjective matters, significantly shapes human decision-making, sometimes leading to uncritical acceptance of AI-generated conclusions.

This research advances our understanding of AI-assisted decision-making in creative evaluation, particularly for tasks involving both subjective and objective criteria. Our findings reveal that human interaction with AI varies significantly based on whether AI provides narratives for its recommendations. We provide early evidence that screeners with AI interaction expertise will more likely scrutinize AI recommendations and integrate them effectively into their decisions. We suggest this form of expertise differs from screeners' domain expertise. Hence, this study lays a foundation for future human-AI collaboration in creative evaluation, highlighting both its potential benefits and challenges.

2. Human-AI Collaboration for Creative Evaluation

Generative AI has shifted the landscape of firm creative problem-solving activities, necessitating a reevaluation of traditional idea generation and evaluation approaches. This transformative technology has made producing new content and ideas remarkably easy, rapid, and cost-effective, leading to a surge in potential solutions across various creative domains (Boussioux et al. 2024, Girotra et al. 2023). This abundance of ideas presents unprecedented opportunities to augment innovation processes. However, it also introduces a significant challenge: the need for effective initial screening to manage the sheer volume of generated content.

To develop a human-AI approach to creative evaluation, we draw on emerging literature that indicates human-AI decision-making can be an effective approach to scale and improve decision-making quality in domains with correct, verifiable responses. This literature focuses on domains with clear rules, patterns, and objectives, such as medical diagnoses (Agarwal et al. 2023, Lebovitz et al. 2022, Tschandl et al. 2020), operations research (Wasserkrug et al. 2024), classification (Wang et al. 2023a), and prediction tasks (Balakrishnan et al. 2022). These areas share a common feature: ground truth outcomes, allowing for clear distinctions between correct and incorrect responses. In such tasks, enhanced performance typically results from combining AI capabilities with human domain expertise, fostering an improved human-AI synergy (Vaccaro et al. 2024).

Creative evaluation, however, is distinct from decision-making in contexts with correct, ground-truth outcomes because it typically entails the consideration of both objective and subjective criteria. Subjective criteria in creative evaluation often involve elements that are inherently difficult to quantify, such as aesthetic appeal, emotional impact, cultural relevance, or originality (Amabile 1983, Boudreau et al. 2016, Elsbach et al. 2003). These aspects often rely on human perception, experience, and cultural context, and whether AI systems can be effective in such domains is unclear. Recent research has begun to examine the use of theory-based models and machine learning for idea screening (Bell et al. 2024) and LLMs for strategic decision-making of business tasks (Csaszar et al. 2024, Doshi et al. 2024), where assessments are based on subjective evaluations of quality. We build on this work to investigate how LLMs can assist creative evaluation. In the remainder of this section, we provide a technical primer on how humans and AI can collaborate to conduct creative evaluations involving objective and subjective criteria. We then investigate the effect of AI rationale on how humans and AI engage in creative evaluation.

2.1. Using LLMs for Creative Evaluation

The emergence of LLMs has opened new horizons in natural language processing capabilities, offering new possibilities for creative evaluation. Built on neural network transformer architectures (Vaswani et al. 2017), LLMs generate text autoregressively, one token (word or subword) at a time, based on probability distributions learned from vast datasets (Gillioz et al. 2020). Unlike rule-based systems or traditional machine learning approaches, LLMs process entire contexts holistically, enabling sophisticated assessments that capture the multifaceted nature of innovative ideas. Their context-aware processing and flexibility allow adaptation to various evaluation criteria without extensive feature engineering and they can leverage knowledge from vast corpora to evaluate ideas across numerous domains (Wei et al. 2022a).

However, they operate on statistical patterns rather than human-like reasoning (Contreras Kallens et al. 2023), which can lead to inconsistent outputs (Bommasani et al. 2021) and risks of users attributing more ‘intelligence’ to LLMs than is actually present (de Véricourt and Gurkan 2023). These limitations are particularly relevant when considering objective and subjective criteria in creative assessments.

Creative evaluation often involves both objective criteria (measurable, fact-based aspects like technical feasibility or market size) and subjective criteria (qualitative judgments like novelty or aesthetic appeal) (Amabile 1983, Highhouse 2008, Kornish and Ulrich 2014). Traditionally, human experts are considered better suited for subjective evaluations. While LLMs can generate assessments that often align with human judgments, even for subjective criteria (Brown et al. 2020, Bender et al. 2021), it is essential to recognize that they are not engaging in logical analysis but in complex pattern matching.

Effective integration of LLMs requires carefully balancing their capabilities with human expertise. This approach should leverage the strengths of both LLMs and human evaluators to navigate the complex interplay between objective and subjective evaluation criteria in creative contexts, potentially enhancing the quality and efficiency of creative idea evaluation while mitigating the limitations of LLMs.

2.2 The Role of AI Narratives in Shaping Human-AI Creative Evaluation

LLM-generated narratives create explanations or rationales for decisions by drawing on patterns and associations learned from extensive training data. These narratives are presented as coherent, contextually relevant text (Ding et al. 2022), making AI decisions more understandable and relatable to users. By providing these rationales, LLMs help bridge the gap between complex AI outputs and human interpretation (Barredo Arrieta et al. 2020), playing a potentially crucial role in AI-assisted idea evaluation.

An intriguing feature of LLMs in creative evaluation is their ability to generate text that appears to “explain” their assessments. However, the assessments and accompanying rationales stem from the same probabilistic text-generation process, rooted in patterns and statistical relationships learned during training

(Bender and Koller 2020). These justifications should be considered plausible narratives aligned with the model's output rather than genuine insights into the AI's reasoning or internal workings. In this regard, the LLM does not properly "reason" to generate recommendations and explanations, and the narratives it provides conceptually differ from the more traditional explainable and interpretable AI outputs (Barredo Arrieta et al. 2020).

In the context of screening, LLM-generated rationales are valuable because they help human evaluators understand and validate the model's assessments, identify potential biases or errors, and refine their own judgments (Ribeiro et al. 2016). This can increase efficiency, as humans can focus on areas where the AI's assessment might be questionable or incomplete. However, generating meaningful rationales for LLM outputs remains challenging due to the complexity of their distributed representations learned from vast amounts of data (Lipton, 2018).

Potential pitfalls in LLM-generated justifications include post-hoc rationalization, inconsistency (Elazar et al. 2021), hallucination risk, sycophancy (i.e., overly positive or agreeable responses to user input) (Sharma et al. 2023), false confidence (Kadavath et al. 2022), misinterpretation of criteria, and reinforcement of biases (Abid et al. 2021). These challenges could be amplified when dealing with subjective criteria, as LLMs base their approximations of human judgment on patterns found in their training data. In such contexts, increased trust in AI does not necessarily lead to improved performance (Ahn et al. 2024). For instance, humans may become overly reliant on AI recommendations by forgoing agency and accountability in the decision-making process (Dell'Acqua 2022, Vasconcelos et al. 2023).

Despite these potential downsides, the unique form of explainability offered by LLM-generated narratives can be valuable for fostering effective human-AI collaboration in idea evaluation. By providing clear and understandable rationales for their recommendations, even if these are contextual justifications rather than true explanations of internal processes, LLM systems can empower human evaluators to critically assess the suggestions' validity and relevance (Ma et al. 2024). This novel approach to AI explainability, distinct from traditional explainable AI methods, warrants further study to understand its potential benefits and limitations in enhancing human-AI collaboration in creative tasks.

3. Research Design and Methodology

3.1. MIT Solve Global Challenges

We partner with an organization within the Massachusetts Institute of Technology (MIT) called MIT Solve. MIT Solve is a marketplace for social impact and social entrepreneurship, connecting startups with funding and resources to solve global challenges. Since its launch in 2015, MIT Solve has received over 23,200 applications from nearly every country and mobilized over $75 million in funding for solver teams and social innovators. Selected applicants receive $10,000 in cash prizes, access to additional funding in the form of grants, investments, and travel stipends, and resources such as workshops, leadership coaching, and networking opportunities during a nine-month personalized support program.

Each year, MIT Solve launches Global Challenges, which address five key areas: Climate, Learning, Health, Economic Prosperity, and Indigenous Communities. The Global Challenges are a series of competitions to find and support innovative, human-centered, technology-based solutions.

MIT Solve uses a rigorous evaluation process to identify winners. Every year, MIT Solve Challenges receives thousands of applications from global innovators, with over 2,200 applicants for the 2024 global challenge competition alone. The selection process consists of several stages. First, all applications are subject to an internal screening process, during which internal staff filter out solutions that do not meet the organization's criteria, which typically corresponds to roughly half of the solutions submitted. All solutions passing the initial screening are considered semi-finalists. These solutions are then sent to external judges in the MIT Solve community, who select the most promising solutions as finalists. Lastly, the finalists pitch their solutions to a panel of leading experts in the field who select the winners.

The primary objective is to ensure that only high-quality solutions advance to the final stages to allocate resources effectively and support the best innovations. To advance this effort, our research focuses on the first stage of the Solve selection process: solution screening. As the number of submissions has steadily increased over the years, the traditional human-centered screening process has exceeded Solve staff's capabilities in terms of resources, time, and effort. These issues highlight the need for enhanced methods that can augment human capabilities and improve the overall quality of the screening process. The motivation for incorporating AI into the screening process is to give additional support to human screeners, improve the quality of the process, and reduce their cognitive burden. Our study focuses on leveraging AI's capabilities to assist screeners in the cognitively demanding process of screening solutions in the first stage of the selection process and to help them decide whether to pass a solution to the second stage, a decision that is based on both objective and subjective criteria.

3.2. Designing the AI Screening Recommendations

MIT Solve uses a systematic process to screen submissions to its global entrepreneurship challenge. The screening process assesses a submission using five criteria, listed in order of priority:

- Criterion 1 - Is the solution application complete, appropriate, and intelligible?
- Criterion 2 - Is the solution at least at the prototype stage?
- Criterion 3 - Does the solution address the challenge question?
- Criterion 4 - Is the solution powered by technology?
- Criterion 5 - Is the quality of the solution good enough that an external reviewer should take the time to read and score it?

Screeners in the human-only process typically evaluate solutions by hierarchically moving down the list of criteria. They screen out a solution if it fails on any criterion. For instance, if a solution fails on criterion 1, the screener immediately fails it and moves on to the next solution. Only solutions passing all criteria advance to the second stage. In this process, two screeners evaluate each solution independently, meeting to reconcile their judgments only in cases of disagreement. Through interviews with our MIT Solve partners and regular solution screeners, we discovered that they consider criteria 1 and 5 subjective, based on opinion or gut instinct. In contrast, they view criteria 2, 3, and 4 as objective, with measurable, factual responses.

In February 2024, we partnered with MIT Solve to design the AI screening recommendations for submissions to the 2024 Global Health Equity Challenge. We used OpenAI's GPT-4, a state-of-the-art LLM, to develop the AI screener model to derive a pass/fail decision for each submission and elicit rationale. To calibrate and refine the decision produced by the LLM for screening, we worked on developing a well-balanced prompt for the models. More specifically, we started by decomposing the screening process into one prompt per selection criterion. Hence, each prompt was intended to verify whether the solution failed the criterion in question. In the next section, we detail the prompting techniques employed to train the LLM to develop recommendations and accompanying rationale on whether to pass or fail a solution, as well as a probability of passing score (expressed as a percentage) for its decision on each criterion. The full text prompt used to generate the recommendations and rationales can be found in Appendix C.

3.3 A Technical Overview Into the "LLM System" for Screening Innovative Submissions

To understand how the LLM evaluates submissions across these diverse criteria, it is important to recognize that the model does not follow a predetermined set of rules or employ discrete logic. Instead, it leverages its vast network of learned associations, processing each submission holistically through its layers of attention mechanisms. We used decisions from MIT Solve's past internal screening decisions for solutions from the 2023 Global Health Challenge to suggest "few-shot" examples of failure and success for each of the five criteria. This process led to generating AI recommendations, accompanied by a short rationale of around 200 words, to pass or fail a submission on each criterion. By including past screening decisions from MIT Solve, we effectively "prime" the model with relevant context. This technique, known as few-shot learning (Brown et al. 2020), helps align the LLM's outputs more closely with the specific evaluation standards of the challenge. For instance, when presented with a new submission alongside examples of previously accepted and rejected solutions, the model can better calibrate its understanding of what constitutes a passing or failing submission in this context.

When evaluating a submission, the LLM simultaneously processes the challenge question, the solution description, and the few-shot examples through its neural networks. This activates various regions in the model's latent space, representing complex, multidimensional features of both the challenge requirements and the proposed solution. The model then generates its evaluation based on the overall pattern of these activations. Consider a challenge seeking "technology to make good health and access to quality care more equitable for all" and a submission proposing "a telehealth platform that uses AI to provide personalized, real-time health consultations and connect patients with local healthcare providers for follow-up care, regardless of their location or socioeconomic status." The LLM does not simply match keywords. Instead, it might recognize that personalized real-time consultations and local healthcare connectivity could improve access to quality care, even if terms like "equity" or "access" are not explicitly mentioned. This recognition stems from the model's learned representations of concepts such as healthcare accessibility, personalized care, and technological integration, all activated in concert to inform the final evaluation.

Moreover, we use chain-of-thought (CoT) mechanisms (Wei et al. 2022b) to improve the model's screening process by forcing it to detail its "reasoning" and layout more precise elements from the solution rather than providing a very general and shallow assessment of the solution being considered.

The model also generates probability scores for pass/fail decisions that reflect the model translating its internal, continuous representations into discrete numerical outputs. This process involves several steps:

1. The model processes the input (challenge question, solution description, and the request for percentages) through its layers, activating relevant patterns in its latent space.
2. The model generates a step-by-step reasoning process output as text. This reasoning is not a trace of actual decision-making but rather an autoregressive probabilistic generation based on the model's training and the input provided.
3. Following the detailed reasoning, the model produces a short explanation summarizing its decision.
4. The model then produces a probability score (as a percentage) for the likelihood of the solution meeting each criterion. This score is generated post-hoc, sampled from a distribution over possible outcomes, influenced by the preceding textual reasoning.

Overall, these outputs are all part of a single, autoregressive generation process, meaning the explanation itself influences the statistical patterns that lead to the final scores and decisions. As a result, the same prompt and solution may yield different evaluations, percentage estimations, and qualitative assessments across multiple runs, even with the same model and input.

3.4 Applying the LLM Recommendations to the 2024 Global Health Equity Challenge

After the submission process for the 2024 Global Health Equity Challenge closed on April 23, 2024, we randomly selected 48 solutions from the 531 submissions for inclusion in the screening experiment (see Figure 2). We applied the pre-trained LLM (detailed above) on the 48 solutions for a total of 240 prompts for the full screening process (5 prompts for each of the 48 solutions in our sample). We set a passing confidence score threshold of 75% for each criterion to effectively calibrate the model's outputs to align with the challenge organizers' objective of screening out more than 50 percent of the initial submissions. However, it is essential to recognize that this confidence score threshold is somewhat arbitrary and that the underlying confidence scores are not true probabilities but artifacts of the model's learned patterns of expressing certainty. This threshold resulted in 20 of 48 (42%) submissions with 'pass' AI recommendations.

3.5 Recruitment of Expert and Community Screeners

Data collection occurred between April 30, 2024, and June 21, 2024. We ran the experiment over four sessions to collect data on human-AI decision-making among domain experts and community screeners. We conducted the first two sessions with domain experts, consisting of 20 internal staff members (MIT Solve staff; session 1) and 52 MIT Solve semi-finalist reviewers and affiliates (MIT Solve judges; session 2). These two sessions produced a total of 72 expert screeners.

We ran the same process for the third and fourth sessions with two types of non-expert community screeners. Session three occurred with 113 undergraduate business students in a Business Data Analytics course for junior and senior students in Information Systems or Operations Management majors at the University of Washington, Foster School of Business (University Students; session 3). Session four took place with 43 students from the Aspire Leaders Program, a global business program for low-income, first-generation college students and recent graduates from around the world, co-founded by two Harvard Business School faculty (Aspire leaders' program; session 4). These two sessions produced a total of 156 community screeners.

3.6 Experimental Design and Study Procedures

We conduct a pre-registered within-subjects experiment with three conditions: a control condition (human-only evaluation), treatment 1 (human-AI collaboration with "black box" AI recommendations), and treatment 2 (human-AI collaboration with "narrative" AI recommendations featuring rationale). In the control condition (C), participants evaluate the submissions without assistance from the generative AI tool, replicating the current evaluation process. In treatment 1 (T1), participants can see the AI's pass/fail recommendations on each criterion to aid their screening decisions. In treatment 2 (T2), participants receive both the AI's recommendations and the rationale behind those recommendations for each criterion. The participants are aware the decision support tool is based on AI.

Each participant was randomly assigned to one of six randomized sequences (e.g., first C then T1, first C then T2, first T1 then T2, etc.), where they screened between 10 and 30 solutions in two of three experimental conditions (5 to 15 screens in each experimental condition). The length of the session determined the number of screened solutions. More specifically, the sessions took place over 60-90 minutes, which included a 12-minute training and overview session conducted by one of the study team's authors. The training session provided an overview of the 2024 Global Health Equity Challenge, an introduction to the screening web application, and the study procedures. We provided compensation to the participants from sessions 1, 2, and 4 (MIT Solve staff, MIT Solve reviewers, and Aspire leaders program students) at the end of the study for their time and contribution, with amounts ranging from $15 to $20 based on the participant group and duration of the study. The undergraduate business students received course credit for completing the study.

Each solution was screened by 62.54 evaluators on average, with around 21 screeners in each condition (see Table B1 in Appendix B for the details). We also conducted balance tests over the main solution level variables (Table B2a in Appendix) and screener level variables (Table B2b in Appendix), suggesting the randomization process mostly achieved balance on observables across the three experimental conditions for both screener and solution characteristics.

The screening process is facilitated through a specially designed interface, as shown in Figure 2, hosted on the Google cloud computing platform. Participants access the interface via a URL provided through Qualtrics. The interface contains the solution details from the application used to make a screening decision (pass or fail) for each submission based on the five predefined criteria. In all three experimental conditions, screeners have the complete solution details for each solution they evaluate on the web application. They also have the opportunity to review criteria definitions and access an AI-generated summary of the solution in a specific tab. Once they have made their final decision, they input on the interface whether they want to pass or fail the solution and select the failure criterion in case of failure. Additionally, screeners input their confidence score for each decision on a Likert scale ranging from 1 to 5, reflecting their certainty in the assessment. This structured interaction ensures that screeners systematically evaluate each submission equivalently regardless of the experimental condition. The interface also features a timer with a 2.5-minute countdown (see Figure 2) to minimize heterogeneity in time and attention allocated to each solution and across screeners. Consistent with the time limit imposed on screeners' decisions, Table A1 indicates there was no difference in decision time by experimental condition.

To familiarize the participants with the web application and the screening process, all participants took a practice quiz consisting of three sample solutions from the 2023 Health in Fragile Contexts Challenge. During the practice quiz, all participants viewed and interacted with the control condition of the web application. Post-screening, the participants completed a short survey to gather additional quantitative and qualitative data on their prior experience with AI decision-making, as well as their perceptions, experiences, and overall satisfaction with the screening process. We also conducted follow-up interviews with a sample of 11 internal staff screeners from MIT Solve to better understand their perceptions and experience with the AI recommendations. The post-screening survey and interview protocol can be found in Appendix D.

Lastly, we collected the mouse movements and time spent on specific application parts using the PostHog system[2] for study sessions 2-4 to determine how AI recommendations altered attention allocation and time management relative to the human-only process.

Figure 2 provides screenshots of the web interface used in the control, black box AI (treatment 1), and narrative AI (treatment 2) conditions.

------------ Figure 2 about here ------------

3.7 Main Variables

Dependent Variables

Our main dependent variable is the screener's *Screening Decision* to pass or fail a solution.

Explanatory Variables

The main explanatory variable is the experimental condition to which the screeners are assigned. We use the variable *AI Treatment Type*, which is a categorical variable that distinguishes between whether the screener was exogenously assigned to the human-only control condition, Treatment 1 (T1), or the black box AI (BBAI) treatment condition with AI recommendations but no rationale, or Treatment 2 (T2), the Narrative AI (NAI) treatment condition, with AI recommendations and rationale.

To determine whether the screeners use the AI recommendations differently depending on the type of AI recommendation, we include the categorical variable *AI Recommendation Type*, corresponding to whether the AI recommended to pass a solution (AI Pass), fail it on objective criteria (AI Objective Fail), or fail it on subjective criteria (AI Subjective Fail). The objective and subjective failure criteria were determined by whether AI recommended failing the solution on one or more objective criteria (criteria 2, 3, 4) or subjective criteria (criteria 1 or 5).

Screener Engagement Behaviors

We assess screeners' engagement with the solution using *Log Mouse Click Engagement*, which is the natural logarithm of the total number of clicks a screener makes on the web interface while interacting with the solution. This metric reflects the screeners' intensity or frequency of interactions with the evaluation interface. Figure A1 illustrates the distribution of screener engagement behaviors as recorded by PostHog.

Other Variables

We also include several variables as covariates in our main analyses investigating screener decisions. First, we use the variable *Domain Expertise* to indicate whether the screener is a domain expert or non-domain expert (1 = non-domain expert, 0 otherwise). Second, we control for confidence, the screener's self-reported confidence in the decision. Confidence is collected on a Likert scale (1 = not at all confident, 5 = highly confident). Third, we control for the sequence order of the experimental condition (1

[2] https://posthog.com/

= second experimental condition in the sequence, 0 otherwise). Lastly, we use solution dummies to account for unobservable differences in the intrinsic quality of the solutions and evaluator dummies to account for unobservable heterogeneity between evaluators. In Table 1, we describe the key variables used in the analyses, as well as their means and standard deviations.

------------ Table 1 about here ------------

Statistical Analyses

We use linear probability models (LPMs) to regress the screeners' decisions on *AI Treatment Type.* Our final model specifications include covariates and solution dummies to account for heterogeneity in the quality of the submitted solutions. We opted for LPMs over logit models for two reasons. First, LPMs provide coefficients directly interpretable as marginal effects, making results more intuitive and easier to communicate. Second, they allow for a straightforward interpretation of interaction terms, which is important for analyzing treatment effects across the subjectivity of the failure criteria. While logit models may offer theoretical advantages for binary outcomes, recent literature suggests that LPMs perform comparably well in practice, especially when probabilities are not extreme (Timoneda 2021). Given these considerations and their simplicity in interpretation, we believe LPMs are well-suited for our analysis.

4. Results

4.1 Experimental Results

We begin by describing the screeners' passing rates and criteria indicated for failing a solution by experimental condition. Table 2 depicts the alignment between the AI screening decisions and the experimental conditions: overall (Table 2a), human-only control condition (Table 2b), BBAI T1 condition (Table 2c), and the NAI T2 condition (Table 2d).

Overall Alignment. Across the full sample, the average alignment between AI decisions and experimental conditions is 67%. This alignment varies by condition: 54% in the human-only control condition, increasing to 73% and 75% in the BBAI T1 and NAI T2 conditions. Interestingly, the alignment between human and AI decisions differs notably depending on whether the AI passes or fails a solution. When AI fails a solution, the alignment is 38% for the human-only condition, up to around 60% and 65% for the BBAI T1 and the NAI T2 treatment conditions, respectively. In contrast, when AI passes a solution, the baseline alignment with AI is already at 78% in the human-only condition, increasing to 88% for both the BBAI T1 and NAI T2 conditions. This suggests that screeners are more likely to align with AI when it passes a solution than when it fails one, with alignment improving further under AI-assisted conditions.

Passing Rates. Table 2 outlines the passing rate of screeners versus AI across experimental conditions. Whereas AI has the lowest passing rate of 42% (i.e., AI recommended passing 20 of 48 solutions), the human-only control condition has the highest passing rate of 68%. The passing rates for the BBAI T1 and NAI T2 conditions are meaningfully lower, at 62% and 57%, respectively. This suggests that humans are more lenient in their evaluations than those in AI-assisted conditions. Interestingly, the NAI T2 condition exhibits a lower passing rate than the BBAI T1 condition, implying that providing rationales for AI decisions might prompt screeners to evaluate solutions more critically or align more closely with AI's recommendations.

------------ Tables 2 about here ------------

Figure 3 shows the solution pass rates across different experimental conditions and AI recommendation types. First, the alignment between human decisions and AI recommendations is notably low when AI recommends failing a solution based on subjective criteria, where alignment drops to 24%. In both AI treatment conditions, we observe an overall increase in alignment compared to the human-only control. Although there is no significant difference in pass rates between BBAI T1 and NAI T2 conditions when AI recommends passing a solution or failing it based on objective criteria, when AI recommends failing a solution based on subjective criteria, the pass rate is lower in the NAI condition compared to the BBAI condition. This suggests more substantial alignment with AI recommendations when rationales are provided, particularly for subjective judgments.

------------ Figure 2 about here ------------

Turning to the regression analyses, Tables 3 and 4 analyze how different types of AI assistance affect screener decisions using LPMs. Table 3 focuses on the overall impact of AI recommendations. We compare how BBAI T1 and NAI T2 recommendations influence screeners' decisions to pass or fail solutions. Table 4 then examines how the type of AI recommendation influences screeners' decisions: AI Pass (Table 4a), AI Objective Fail (Table 4b), and AI Subjective Fail (Table 4c). Specifically, we investigate whether providing rationale changes how screeners respond to subjective versus objective AI recommendations for failure.

Turning to Table 3, in Model 1, we find that screeners are significantly more likely to fail a solution in the AI treatment conditions than in the human-only control condition (Model 1: -0.090, $p < 0.01$). In Model 2, we differentiate between the two treatment conditions and show that the likelihood of failing a solution is higher in the NAI T2 than in the BBAI T1 condition (BBAI: -0.065, $p < 0.01$; NAI: -0.115, $p < 0.01$), compared to the human-only control condition. Notably, an ANOVA test of Model 2's coefficients indicates that they are statistically different from each other ($F(2,2999) = 14.496$, $p < 0.001$), suggesting a more pronounced influence of NAI T2. These effects remain robust across Models 3 and 4, incorporating additional covariates and solution-specific controls. Intriguingly, Model 5 indicates that these AI effects are consistent across domain experts and non-experts, with no significant interaction observed for either the BBAI T1 or NAI T2 conditions. This suggests that the influence of AI recommendations transcends individual expertise levels, even though we observe in Figure A2 that the domain experts are more consistent in their screening decisions in both the human-only control and BBAI T1 conditions. Lastly, Model 6 demonstrates that the reported coefficients in Table 3 Models 1-4 remain robust and significant after including the evaluator dummies to account for evaluator-specific heterogeneities. These findings underscore the substantial impact of AI, particularly NAI T2, on human decision-making in screening tasks—highlighting the potential of AI to systematically alter decision outcomes, particularly when humans and AI arrive at different decisions.

------------ Table 3 about here ------------

Table 4 investigates how the experimental conditions influence the screeners' decisions by *AI Recommendation Type* to gain insight into how screeners' behaviors differed depending on the valence (pass vs. fail) and criteria for failure (objective vs. subjective).

AI Recommends Pass. Table 4a indicates how the screeners' decisions are influenced by the AI's recommendation to pass a solution. First, the constant term in Model 1 of 0.780 indicates a relatively high level of alignment in perceptions between the AI's recommendation to pass a solution and the human-only control condition. Next, Model 1 demonstrates that screeners are 10.4 pp or 13% more likely to pass a solution when AI recommends passing a solution in both the BBAI T1 (Model 1: 0.104, $p < 0.01$) and NAI T2 condition (Model 1: 0.104, $p < 0.01$), compared to the no AI assistance control condition. These effects remain robust to adding covariates and solution dummies in Models 2 and 3. In Model 4, we introduce the interaction term between the experimental conditions and domain expertise. Here, we observe that the treatment effect is driven by non-experts in the NAI T2 condition (Model 4: 0.135, $p < 0.01$), but there is no difference between experts and non-experts in the BBAI T1 condition (Model 4: 0.061, *ns*). Lastly, we show in Model 5 that the reported effects are robust to screener dummies in the BBAI T1 (Model 5: 0.113, $p < 0.01$) and NAI T2 conditions (Model 5: 0.142, $p < 0.01$).

AI Recommends Objective Failure. Table 4b investigates how the screeners' decisions are influenced by the AI's recommendation to fail a solution based on objective criteria. The constant term in Model 1 of 0.407 indicates a reasonable level of concordance of 59% between the human-only control and AI's objective failure recommendations. Next, Model 1 indicates that screeners are 22.9 pp (or 56.3%) and 19.6 pp (or 48.2%) more likely to fail a solution when AI suggests failing it on objective criteria. Pairwise comparisons between the levels of the *AI Treatment* variable indicate no difference between the BBAI T1 and NAI T2 conditions on screener behaviors ($p = 0.689$). The coefficients for the AI treatments remain stable in Models 2 and 3, which add covariates and solution dummies. In Model 4, we observe that the AI treatment effects are similar for experts and non-experts. Finally, Model 5 indicates that the reported effects in Models 1-3 are robust to the addition of screener dummies in the NAI treatment condition (Model 5: -0.155, $p < 0.05$) and directionally robust in the BBAI treatment condition (Model 5: -0.104, *ns*).

AI Recommends Subjective Failure. Table 4c investigates how the screeners' decisions are influenced by the AI's recommendation to fail a solution based on subjective criteria. First, the constant term in Model 1 indicates that the pass rate for the human-only condition is 0.763 when the AI recommends failing based on subjective criteria, suggesting a high degree of misalignment in perceptions. In particular, the human-only condition is only 24% aligned with AI on failures based on subjective criteria. Examining the AI treatment effects, we observe in Model 1 that screeners are 20.3 pp (or 26.6%) in the BBAI T1 condition and 31.8 pp (or 41.7%) in the NAI T2 condition more likely to fail a solution than the human-only control condition. A pairwise comparison of the coefficients for the BBAI T1 and NAI T2 conditions indicates that they are statistically different (p = 0.006), with the NAI T2 condition having a more substantial effect on screening behaviors than the BBAI T1 condition. Models 2 and 3 indicate that the reported effects are stable after adding covariates and solution dummies. Once again, Model 4 shows no difference in treatment effects by domain expertise. Lastly, Model 5 indicates that the reported effects in Models 1-3 are stable to the inclusion of screener dummies in both the BBAI T1 (Model 5: -0.197, $p < 0.01$) and NAI T2 (Model 5: -0.312, $p < 0.01$) conditions.

------------ Tables 4 about here ------------

4.2 Screener Engagement Results

In this section, we examine how screeners' decisions vary across experimental conditions based on their level of engagement, as measured by mouse clicks on the web application. The analysis utilizes data from sessions 2-4, during which data was collected using the PostHog system. Figure A1 illustrates a marginally significant difference in the *Log mouse clicks* across experimental conditions (F(2,1956) = 2.406, p = 0.090).

Table 5, Model 1 indicates that with each one-unit increase in log mouse clicks, the log odds of passing a solution decrease by 0.203 units, equivalent to an 18% reduction in the odds of passing (exp(-0.203) = 0.816). This coefficient suggests that, on average, participants are more likely to fail solutions they engage with more. Model 2 incorporates interaction terms between the experimental conditions and *Log mouse clicks*. These interaction terms demonstrate that the effects of treatment conditions are consistent across different levels of screener engagement.

Models 3-5 provide further analysis by breaking down screener behaviors according to AI recommendation types. When the AI recommends passing a solution (Model 3), the interaction terms between the treatment conditions and *Log mouse clicks* are not significant, suggesting that screener engagement does not significantly differ by the presence of an AI recommendation. However, Models 4 and 5 reveal contrasting behaviors when AI uses objective and subjective criteria to recommend failing a solution. In Model 4, for AI objective failures, the interaction term between *NAI T2* x *Log mouse clicks* is positive and significant (Model 4: 0.285, $p < 0.05$). This coefficient suggests that a one-unit increase in *Log* *mouse clicks* increases the odds of passing a solution by 33% under the NAI treatment (exp(0.285) = 1.330). The interaction term between *BBAI T1* x *Log mouse clicks* is positive but not significant (Model 4: 0.174, *ns*). This coefficient suggests that participants in the NAI T2 condition are more likely to override AI's failing recommendation on objective criteria when they engage more intensely with the solution.

Model 5 presents an opposite trend for AI subjective failures. Here, the interaction term *BBAI T1* x *Log mouse clicks* is negative and significant (Model 5: -0.211, $p < 0.05$), indicating that each one-unit increase in *Log mouse clicks* reduces the odds of passing by 19% (exp(-0.211) = 0.810). This suggests that AI narratives may alter how screeners interact with the solution's content, having different implications depending on the recommendation's subjectivity level.

The differences in engagement patterns and screening decisions observed in Models 4 and 5 may stem from the varying degrees of baseline alignment between human decisions and AI recommendations for objective versus subjective criteria. When the AI recommends failing a solution based on objective criteria, screeners in the NAI treatment are more likely to override these recommendations as their interaction with the solution increases. This suggests that screeners may disagree with the AI's rationale on objective failures and thus engage more deeply to gather the necessary information to justify passing the solution. These behaviors are consistent with the moderate alignment (59% agreement: see Table 4b, Model 1) between the human-only control and AI's recommendations on objective criteria. In contrast, when the AI recommends failure based on subjective criteria, screeners' deeper engagement with the content in the BBAI treatment may indicate their effort to thoroughly evaluate whether the AI's subjective assessment is valid and whether the solution should indeed fail. These behaviors reflect the low degree of alignment (24% agreement: see Table 4c, Model 1) between the human-only control condition and the AI's recommendations for subjective failures. In the BBAI (T1) condition, participants may engage more deeply with the solutions because they need to scrutinize why the BBAI suggested failing based on subjective criteria, prompting them to explore further. In contrast, in the NAI (T2) condition, participants are provided with a rationale that explains the AI's failure on subjective criteria. This narrative reduces the need for additional engagement, leading participants to rely on the explanation and more readily decide to defer to AI's recommendations.

------------ Table 5 about here ------------

5. Discussion and Conclusion

This study explores the transformative potential of generative AI, particularly LLMs, in enhancing creative idea evaluation through human-AI collaboration, focusing on the interplay between objective and subjective criteria. Using real-world solutions submitted to MIT Solve's 2024 Global Health Equity Challenge, our findings reveal complex dynamics in how screeners interact with AI recommendations, particularly when evaluating subjective aspects of innovative solutions. We observe that screeners were more likely to fail solutions with AI assistance, especially based on subjective criteria. Screeners were significantly more likely to adhere to AI recommendations for subjective criteria when provided with a rationale, despite the AI's lack of true understanding of the fact-opinion distinction. Notably, while these effects were consistent across both domain experts and non-experts, evaluators who engaged more deeply with the solutions were more likely to critically scrutinize the LLM's recommendations. This observation suggests the emergence of a new form of expertise—AI interaction expertise—which involves effectively interpreting, questioning, and integrating AI-generated insights into decision-making processes. [3]

Lastly, in supplementary analyses, we compare the performance of the screeners' decisions in the three experimental conditions on various quality dimensions, as rated by external judges (see Table A3). Our analysis indicates that screeners with access to BBAI recommendations outperformed the human-only control group and the NAI treatment condition in selecting higher-quality solutions. This suggests that while AI rationale can aid screeners with understanding AI's recommendations, they may not always lead to better outcomes regarding solution quality.

5.1. Balancing Objectivity and Subjectivity: AI's Role in Creative Evaluation

Our findings highlight several key differences between purely objective evaluations and those that intermix subjectivity and objectivity, particularly in how humans interpret and rely on AI recommendations in different contexts. We demonstrate that human decision-makers and AI recommendations may differ in their degree of alignment, depending on the valence and the subjectivity of the decision. Whereas the human-only control and AI had relatively high alignment among solutions for which AI recommended passing a solution and moderate alignment among solutions for which AI recommended failing on objective criteria, the alignment was poor for subjective decisions. Due to the correlated nature of the human screeners and AI's recommendations for passing solutions, this suggests that there is a greater opportunity for human-AI collaboration when AI recommends failing a solution.

Our study's findings suggest that the implications of AI assistance may differ depending on whether AI provides objective or subjective assessments. In particular, the 59% baseline agreement for objective criteria indicates a reasonable level of concordance between human judgments and AI recommendations. With AI support, the level of agreement in the experimental conditions increased to around 80%. For objective criteria, our study's findings suggest that AI can offer human screeners additional support for their decision-making. Since AI systems tend to perform well with quantifiable, measurable data (and can be further enhanced by incorporating more high-quality objective data), our findings point to the potential for using LLMs to pre-screen or filter decisions based on objective criteria before passing them onto human

[3] The higher scrutiny leading to an overriding of AI recommendation is further confirmed by the higher alignment of individuals trusting AI with its recommendations. Their lack of AI expertise to question and scrutinize AI recommendations is suggested by the positive coefficient of the interaction between AI Trust and the two treatments on screening decision in Table A3 in Appendix A.

screeners. This first layer of the LLM screener can filter out cases that do not meet objective criteria to reduce the cognitive load on human screeners.

The low baseline agreement of 24% between human screeners and AI recommendations for subjective criteria reveals a significant gap between human reasoning and AI assessments. Our interviews with screeners highlighted two potential factors. First, humans often give solutions the benefit of the doubt, particularly when no clear objective reason for rejection exists—a nuance that AI, focused on strict criteria, may not recognize or account for. However, this human tendency may not be sustainable in the long run, given the organization's limited capacity to advance solutions to the next stage of the application process. Another factor is that the LLM may not be fully attuned to the nuances of the problem domain, possibly due to inadequate training data, model complexity, or the nature of the task. Although alignment with AI's recommendations was higher in the treatment conditions for subjective assessments, our findings suggest that effective decision-making for subjective criteria requires human oversight and close collaboration with AI.

5.2. Designing Effective Human-AI Collaboration with AI-Generated Narratives

AI-generated narratives introduce a new dimension to the evaluation process. They may provide a valuable basis for dialogue between human evaluators and AI systems but also influence human decision-making through personalized persuasive messages that significantly impact attitudes and behavioral intentions (Chu and Liu 2023, Matz et al. 2024). This interaction allows for a human-in-the-loop approach, positioning the LLM as more of a prescriptive model than a purely predictive one. The differences or discrepancies between AI-generated rationales and human judgments can catalyze deeper analysis and discussion. When human evaluators encounter AI narratives that differ from their initial assessments, it prompts them to reconsider their viewpoints, examine the AI's reasoning, and engage in more thorough deliberation. This process of reconciling differing perspectives can lead to more nuanced and carefully considered evaluations, potentially improving the overall quality of decision-making.

Our study underscores the complex dynamics when human evaluators interact with AI-generated recommendations and narratives "explaining" their decisions. The results particularly challenge the assumption that more justification always leads to better outcomes in AI-assisted decision-making, calling for a reevaluation of how and when AI rationale should be provided in collaborative settings (Miller 2019, Doshi-Velez and Kim 2017). Building on our study's insights, we propose several key considerations for designing effective human-AI collaboration systems for creative evaluation. First, future systems can seek to implement differentiated rationale strategies, tailoring AI outputs to the nature of the criteria being evaluated. AI could provide a more direct, data-driven rationale for objective criteria based on its training data and references to the solution text. In contrast, for subjective criteria, AI can offer more tentative, exploratory insights, perhaps presenting multiple perspectives or highlighting the nuances of the evaluation.

This approach aligns with our finding that screeners were more likely to adhere to AI recommendations for subjective criteria when provided with a rationale, suggesting a need for careful framing of these more interpretive assessments.

Second, the nuanced interrelationship between screeners' level of engagement with AI recommendations observed in our study underscores the need for more dynamic, interactive interfaces. Our findings suggest that AI interaction expertise—the ability to effectively interpret and critically assess AI-generated insights, plays a crucial role in how evaluators respond to AI rationales. In particular, our study showed that the impact of AI-generated rationale can vary depending on their capacity to scrutinize and, when necessary, challenge the AI's conclusions. This heterogeneity in response underscores the importance of tailoring the implementation of LLMs to the specific context and user base of each evaluation system. This is consistent with research showing that worker experience and seniority (Wang et al. 2023a) and nature and difficulty of the task (Bayer and Renou 2024) can influence the effectiveness of human-AI collaboration.

To accommodate differences in engagement, these interfaces can be designed to position AI as a 'sounding board' rather than a 'ghost evaluator' (Chen and Chan 2023). Similarly to how companies have to develop agile strategies for integrating enterprise systems (Aral et al. 2024), choosing the right AI implementation strategy for creative evaluations can significantly impact the benefits realized. In a 'sounding board' approach, evaluators would actively engage with the AI system, probing its assessments, seeking clarifications, and exploring alternative perspectives on specific aspects of an idea. This design encourages evaluators to use AI recommendations as a catalyst for their own critical thinking rather than passively accepting them. The 'Sounding Board' concept is particularly crucial for evaluating subjective criteria, where our findings indicate that screeners were more susceptible to AI influence in the NAI treatment condition. Evaluators can leverage AI insights by fostering a more dialogic interaction while maintaining their decision-making autonomy. In follow-up interviews conducted after the study, several screeners suggested improvements to the interface design that would facilitate this 'sounding board' approach. These suggestions often centered on allowing evaluators to formulate their initial opinions before exposure to AI recommendations, thereby enhancing critical thinking and mitigating potential biases from AI input, especially for subjective criteria. This approach would not only preserve the value of human judgment but also maximize the potential of AI as a complementary resource in the creative evaluation process. However, this method may have a drawback: it could require screeners to invest more time and mental effort in each evaluation.

Third, the varying levels of adherence to the AI recommendations suggest the importance of comprehensive training for human evaluators. Such training should emphasize how to critically engage with AI rationale and understand their value and limitations. Organizations can develop extensive training programs and guidelines for individuals working with AI-assisted evaluation systems, emphasizing the importance of critically assessing AI recommendations, especially for subjective criteria (Araujo et al. 2020, Logg et al. 2019). A critical element of these training programs is to highlight the fundamental differences between human and AI reasoning, particularly in the context of subjective assessments, to prevent the kind of anthropomorphizing observed in some of our interview responses. By implementing these strategies, organizations can work towards creating more effective human-AI collaboration systems for creative evaluation. These systems would leverage the strengths of both human insight and AI capabilities while mitigating the potential pitfalls highlighted by our research.

We can create more transparent and trustworthy AI-assisted idea-screening systems by acknowledging the limitations of LLM-generated rationale, particularly in subjective domains, and developing thoughtful approaches to their use and interpretation. This nuanced integration of AI into the creative evaluation process respects the complexity of subjective assessments while harnessing the analytical power of AI for objective criteria. As we continue to explore the potential of LLMs in creative evaluation, this balanced approach will be essential in maximizing the benefits of AI assistance while maintaining the irreplaceable value of human judgment in identifying and nurturing innovative ideas.

5.3. Strategic Implications for Human-AI Collaboration in Creative Evaluation

Our study illustrates that LLMs can strategically impact human decision-making in creative evaluation. Two strengths of LLM-based systems are their adaptability and versatility. These models can be calibrated through techniques like few-shot learning to align with an organization’s specific evaluation criteria, historical data, and desired level of stringency. This adaptability allows for more conservative or lenient evaluation models based on the organization’s goals. For example, suppose past data indicates that promising innovations were occasionally overlooked due to overly strict initial screening. In that case, the LLM can be prompt-tuned to be more inclusive in its early-stage evaluations. The versatility of LLMs also enables the development of multi-stage screening processes. Organizations can design cascading evaluation systems where LLMs make initial broad assessments based on objective criteria, followed by more focused, criteria-specific evaluations. This approach allows for gradually refining the solution pool, potentially reducing the workload on human evaluators while maintaining a high selection standard. Moreover, as organizations evolve their values, criteria, or challenge focus, LLMs can be promptly reconfigured to reflect these changes, offering agility in dynamic fields where evaluation standards may shift rapidly.

Beyond their role in direct evaluation, LLMs can serve as powerful tools for stimulating human cognitive processes (Boyacı et al. 2024). By providing detailed rationales for their assessments, LLMs can offer second opinions that challenge human evaluators’ initial judgments, possibly encouraging deeper reflection and more thorough analysis. This interaction between AI-generated insights and human expertise can lead to more robust and well-considered evaluations.

It is crucial to recognize that ground truth is often elusive, subjective, and challenging to define in many creative evaluation contexts. Implementing LLMs allows one to build fairer evaluation systems by reducing certain human biases and inconsistencies. However, this potential for increased fairness must be balanced against the risk of introducing new algorithmic biases. The key lies in leveraging LLMs as tools to augment human decision-making rather than replace it entirely.

5.4. Limitations and Future Directions

While our study provides valuable insights, it has several limitations that point to directions for future research. Our focus on a specific context raises questions about generalizability. In this scenario, we established a setting where the organization's resources and processes, as reflected in the AI's recommendations, applied more stringent advancement criteria than the human screeners. This created some degree of misalignment between human and AI decisions. The human screeners operated with more optimistic priors about each solution's potential than the organization's official stance. These aspects of our context are likely to influence how humans interact with AI recommendations, compared to cases where human and AI decisions might be more correlated (Agarwal et al. 2023) or more divergent (Balakrishnan et al. 2022). Because organizational objectives and priorities influence the design of AI systems, future research can explore whether these findings hold in other domains of creative evaluation, extending the work on human-AI collaboration to other creative contexts.

Our experimental design balanced the need for consistency and efficiency in the AI's recommendations and the complexity of implementation to improve user engagement and contextual adaptation. We adopted pre-trained, static AI recommendations to improve our interventions' consistency, efficiency, and scalability. Because all users receive the same recommendations for a given solution, this allowed us to vet and validate the recommendations before deployment and ensure the uniformity of the AI decision support. However, this meant that our AI recommendations could not adjust to user-specific queries, and screeners could not explore alternatives or additional rationale. While interactive recommendations would have offered more flexibility, active participation, and critical thinking, a tradeoff to consider may be the consistency and quality of the AI recommendations. In reality, organizations may adopt a hybrid approach that could begin with either no recommendation or static recommendations as a baseline with the option for interactive follow-up queries that would allow for more customized and complex decision-making processes (Ehsan et al. 2021, Liao et al. 2020). Such system designs will likely encourage critical engagement with AI recommendations to reduce overreliance tendencies.

Future systems could implement ensemble methods to address the potential biases and limitations of single LLM evaluations (Doshi et al. 2024). This approach would combine multiple LLM evaluations using different prompting strategies, underlying models, or training data. It could also include running the same creative evaluation multiple times to provide a more comprehensive and nuanced assessment of each idea’s strengths and weaknesses. This strategy, analogous to the “wisdom of the crowd” concept, would better reflect the diverse and potentially conflicting opinions in the models’ training data, capturing different viewpoints or interpretations. Importantly, ensemble methods could help mitigate the risk of over-reliance on a single AI perspective, a concern highlighted by our finding that screeners were significantly more likely to adhere to AI recommendations for subjective criteria when rationale was provided.

Our study reveals the complex dynamics of human-AI collaboration in creative evaluation. While AI assistance can significantly influence screening decisions, the relationship between AI rationale and decision quality is complex. As organizations increasingly integrate AI into evaluation processes, it is essential to carefully design these systems to leverage AI’s strengths while preserving human critical thinking and expertise. We believe this paper can lay the foundation for additional studies to explore the role of human-AI collaboration in creative evaluation involving decision-making based on objective and subjective criteria.

6. References

Abid A, Farooqi M, Zou J (2021) Persistent anti-Muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

Agarwal N, Moehring A, Rajpurkar P, Salz T (2023) Combining human expertise with artificial intelligence: Experimental evidence from radiology. *National Bureau of Economic Research*.

Ahn D, Almaatouq A, Gulabani M, Hosanagar K (2024) Impact of Model Interpretability and Outcome Feedback on Trust in AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-25).

Allen R, Choudhury P (2022) Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organ. Sci.* 33(1):149–169.

Amabile TM (1983) The social psychology of creativity: A componential conceptualization. *J. Pers. Soc. Psychol.* 45(2):357.

Aral S, Brynjolfsson E, Gu C, Wang H, Wu DJ (2024) Understanding the returns from integrated enterprise systems: the impacts of agile and phased implementation strategies. *MIS Quarterly*, *48*(2).

Araujo T, Helberger N, Kruikemeier S, de Vreese CH (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society* 35(3):611–623.

Balakrishnan M, Ferreira K, Tong J (2022) Improving human-algorithm collaboration: Causes and mitigation of over-and under-adherence. *Available SSRN 4298669*.

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58:82–115.

Bayer R C, Renou L (2024) Interacting with Man or Machine: When Do Humans Reason Better? *Manag. Sci.*

Bayer S, Gimpel H, Markgraf M (2022) The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, *32*(1), 110-138.

Bell JJ, Pescher C, Tellis GJ, Füller J (2024) Can AI help in ideation? A theory-based model for idea screening in crowdsourcing contests. *Mark. Sci.* 43(1):54–72.

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bender E, Koller A (2020) Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Bohnet I, Van Geen A, Bazerman M (2015) When performance trumps gender bias: Joint vs. separate evaluation. *Manag. Sci.* 62(5):1225–1234.

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, ... Liang P (2021) On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Boudreau KJ, Guinan EC, Lakhani KR, Riedl C (2016) Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Manag. Sci.* 62(10):2765–2783.

Boussioux L, N Lane J, Zhang M, Jacimovic V, Lakhani KR (2024) The Crowdless Future? Generative AI and Creative Problem-Solving. *Organ. Sci.* 0(0).

Boyacı T, Canyakmaz C, de Véricourt F (2024) Human and machine: The impact of machine input on decision making under cognitive limitations. *Manag. Sci.*, 70(2), 1258-1275.

Brand J, Israeli A, Ngwe D (2023) Using GPT for market research. *Harv. Bus. Sch. Mark. Unit Work. Pap.* (23-062).

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. (2020) Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165.

Brynjolfsson E, Li D, Raymond LR (2023) *Generative AI at work* (National Bureau of Economic Research).

Burtch G, Lee D, Chen Z. (2024) The consequences of generative AI for online knowledge communities. *Scientific Reports*, 14(1), 10413.

Burton JW, Stein MK, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making*, *33*(2), 220-239.

Castelo N, Bos MW, Lehmann DR (2019) Task-Dependent Algorithm Aversion. *J. Mark. Res.* 56(5):809–825.

Chai S, Doshi AR, Silvestri L (2021) How Catastrophic Innovation Failure Affects Organizational and Industry Legitimacy: The 2014 Virgin Galactic Test Flight Crash. *Organ. Sci.*

Chen Z, Chan J (2023) Large language model in creative work: The role of collaboration modality and user expertise. *Available SSRN* 4575598.

Chen V, Liao QV, Wortman Vaughan J, Bansal G (2023) Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.* 7(CSCW2):1–32.

Chu H, Liu S (2023) Can AI tell good stories? Narrative Transportation and Persuasion with ChatGPT.

Contreras Kallens P, Kristensen-McLachlan RD, Christiansen MH (2023) Large language models demonstrate the potential of statistical learning in language. *Cogn. Sci.* 47(3):e13256.

Criscuolo P, Dahlander L, Grohsjean T, Salter A (2017) Evaluating novelty: The role of panels in the selection of R&D projects. *Acad. Manage. J.* 60(2):433–460.

Csaszar FA, Eggers JP (2013) Organizational decision making: An information aggregation view. *Manag. Sci.* 59(10):2257–2277.

Csaszar FA, Ketkar H, Kim H (2024) Artificial Intelligence and Strategic Decision-Making: Evidence from Entrepreneurs and Investors. *arXiv preprint arXiv:2408.08811*.

Csikszentmihalyi M (1999) 16 implications of a systems perspective for the study of creativity. *Handb. Creat.* 313.

Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint* arXiv:2006.11371.

Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR (2023) Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harv. Bus. Sch. Technol. Oper. Mgt Unit Work. Pap.* (24–013).

Dell'Acqua F (2022) Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters. Work. Pap.

Ding W, Abdel-Basset M, Hawash H, Ali AM (2022) Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, *615*, 238-292.

Doshi AR, Bell JJ, Mirzayev E, Vanneste B (2024) Generative Artificial Intelligence and Evaluating Strategic Decisions. *Available at SSRN.*

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608.

Ehsan U, Liao QV, Muller M, Riedl MO, Weisz JD (2021) Expanding explainability: Towards social transparency in AI systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

Elazar Y, Kassner N, Ravfogel S, Ravichander A, Hovy E, Schütze H, Goldberg Y (2021) Measuring and Improving Consistency in Pretrained Language Models Roark B, Nenkova A, eds. *Trans. Assoc. Comput. Linguist.* 9:1012–1031.

Elsbach KD, Kramer RM (2003) Assessing creativity in Hollywood pitch meetings: Evidence for a dual-process model of creativity judgments. *Acad. Manage. J.* 46(3):283–301.

Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Coughlin JF, Guttag JV, Colak E, Ghassemi M (2021) Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* 4(1):31.

Gillioz A, Casas J, Mugellini E, Abou Khaled O (2020) Overview of the transformer-based models for NLP tasks. *2020 15th Conference on Computer Science and Information Systems* (FedCSIS), 179–183.

Girotra K, Meincke L, Terwiesch C, Ulrich KT (2023) Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available SSRN 4526071*.

Gompers P, Gornall W, Kaplan SN, Strebulaev IA (2021) How Venture Capitalists Make Decisions An inside look at an opaque process. *Harvard Business Review*, 99(2), 70-+.

Hammedi W, van Riel AC, Sasovova Z (2011) Antecedents and consequences of reflexivity in new product idea screening. *Journal of Product Innovation Management*, *28*(5), 662-679.

Highhouse S (2008) Stubborn reliance on intuition and subjectivity in employee selection. *Ind. Organ. Psychol.* 1(3):333–342.

Huang H, Fu R, Ghose A (2023) Generative AI and Content Creators: Evidence from Digital Art Platforms *Available at SSRN* 4670714.

Huang L, Pearce JL (2015) Managing the Unknowable: The Effectiveness of Early-stage Investor Gut Feel in Entrepreneurial Investment Decisions. *Administrative Science Quarterly*, 60(4), 634-670.

Jacobs M, Pradier MF, McCoy Jr TH, Perlis RH, Doshi-Velez F, Gajos KZ (2021) How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl. Psychiatry* 11(1):108.

Jia N, Luo X, Fang Z, Liao C (2023) When and how artificial intelligence augments employee creativity. *Acad. Manage. J.* (ja).

Kadavath S, Conerly T, Askell A, Henighan T, Drain D, Perez E, Schiefer N, et al. (2022) Language Models (Mostly) Know What They Know. (November 21) http://arxiv.org/abs/2207.05221.

Kim M, DellaPosta D (2022) The fickle crowd: Reinforcement and contradiction of quality evaluations in cultural markets. *Organ. Sci.*, *33*(6), 2496-2518.

Kornish LJ, Ulrich KT (2014) The importance of the raw idea in innovation: Testing the sow's ear hypothesis. *J. Mark. Res.* 51(1):14–26.

Krakowski S, Luger J, Raisch S (2023) Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal*, *44*(6), 1425-1452.

Lane J, Szajnfarber Z, Crusan J, Menietti M, Lakhani KR (2023) When Does Project Feasibility Drive Technological Innovation? Evaluator Expertise Range, Architectural Knowledge, and Preferences for Existing Technologies.

Lebovitz S, Lifshitz-Assaf H, Levina N (2022) To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organ. Sci.* 33(1):126–148.

Li D (2017) Expertise versus bias in evaluation: Evidence from the NIH. *American Economic Journal*: *Applied Economics* 9(2):60–92.

Liang W, Zhang Y, Wu Z, Lepp H, Ji W, Zhao X, Cao H, Liu S, He S, Huang Z, Yang D (2024a) Mapping the increasing use of LLMs in scientific papers. *arXiv preprint arXiv:2404.01268*.

Liang W, Zhang Y, Cao H, Wang B, Ding DY, Yang X, Vodrahalli K, He S, Smith DS, Yin Y, McFarland DA (2024b). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, *1*(8), AIoa2400196.

Liao QV, Gruen D, Miller S (2020) Questioning the AI: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3):31–57.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. 151:90–103.

Lou B, Wu L (2021) AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms. *Manag. Inf. Syst. Q.* 45(3):1451–1482.

Ma S, Zhang C, Wang X, Ma X, Yin M (2024) Beyond Recommender: An Exploratory Study of the Effects of Different AI Roles in AI-Assisted Decision Making. *arXiv preprint* arXiv:2403.01791.

Matz SC, Teeny JD, Vaid SS, Peters H, Harari GM, Cerf M (2024) The potential of generative AI for personalized persuasion at scale. *Scientific Reports* 14(1):4692.

Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267:1–38.

Mollick E, Nanda R (2016) Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Manag. Sci.* 62(6):1533–1553.

Piezunka, H., & Dahlander, L. (2015). Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. *Academy of Management Journal*, 58(3), 856–880.

Raisch S, Fomina K (2023) Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Acad. Manage. Rev.* (ja):amr. 2021.0421.

Ribeiro MT, Singh S, Guestrin C (2016) " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, Cheng N, et al. (2023) Towards Understanding Sycophancy in Language Models. (October 27) http://arxiv.org/abs/2310.13548.

Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, Brown AR, et al. (2023) Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint* http://arxiv.org/abs/2206.04615.

Timoneda JC (2021) Estimating group fixed effects in panel data with a binary dependent variable: How the LPM outperforms logistic regression in rare events data. Social Science Research 93:102486.

Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, ... & Kittler H (2020) Human–computer collaboration for skin cancer recognition. *Nature medicine* 26(8):1229–1234.

Vaccaro M, Almaatouq A, Malone T (2024) When Are Combinations of Humans and AI Useful? *arXiv preprint* http://arxiv.org/abs/2405.06087.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Advances in neural information processing systems* 30.

Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R (2023) Explanations can reduce overreliance on AI systems during decision-making. *Proc. ACM Hum.-Comput. Interact.* 7(CSCW1):1–38.

de Véricourt F, Gurkan H (2023) Is your machine better than you? You may never know. *Manag. Sci.*

Wang W, Gao G, Agarwal R (2023a) Friend or foe? Teaming between artificial intelligence and workers with variation in experience. *Manag. Sci.,* 70(9), 5753-5775.

Wang W, Yang M, Sun T (2023b) Human-AI co-creation in product ideation: The dual view of quality and diversity. *Available at SSRN 4668241.*

Wasserkrug S, Boussioux L, Hertog DD, Mirzazadeh F, Birbil I, Kurtz J, Maragno D (2024) From large language models and optimization to decision optimization copilot: A Research Manifesto. *arXiv preprint* arXiv:2402.16269.

Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, ..., Fedus W (2022a) Emergent abilities of large language models. *arXiv preprint* arXiv:2206.07682.

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D (2022b) Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.

Wu L, Lou B, Hitt LM (2024) Innovation Strategy After IPO: How AI Analytics Spurs Innovation After IPO. *Manag. Sci.*

Zhou E, Lee D (2024) Generative AI, human creativity, and art. *PNAS Nexus*, 3(3), 052.

Tables and Figures

Table 1. Summary Statistics and Description of Main Variables

Variable Name	Description	Mean (s.d.)
	Dependent Variables	
Screener Decision	A dummy variable = 1 if a screener passes a solution and 0 otherwise	0.625 (0.484)
AI Decision Override	A dummy variable = 1 if a screener chooses to override the AI's decision/recommendation	0.331 (0.471)
	Explanatory Variables	
AI Treatment Type	A categorical variable with three levels: Human-only control (C), Black box AI (T1), and Narrative AI (T2)	1.97 (0.820)
AI Recommendation Type	A categorical variable with three levels: AI passes the solution (41.6%), AI fails it based on objective criteria (22.9%), AI fails it based on subjective criteria (35.4%)	0.918 (0.868)
	Covariates	
Domain expertise	A dummy variable = 1 if the screener is a non-expert and 0 otherwise	0.577 (0.494)
Confidence	The confidence (1 = not at all confident, 5 = highly confident) the screener indicated in their decision	3.20 (1.63)
Sequence	A dummy variable = 1 if the solution corresponds to the second experimental condition and 0 otherwise	0.495 (0.500)

Table 2a: Total - AI Pass vs. Screener Pass in all Experimental Conditions

Screener Pass	**AI Pass = 0**	**AI Pass = 1**	**Total**
0	31%	6%	37%
1	27%	36%	63%
Total	58%	42%	100%

Alignment in all Experimental conditions: 67% (When AI fails: 54%; When AI passes: 85%)

Table 2b: Control - AI Pass vs. Screener Pass in Human-only Control Condition

Screener Pass	**AI Pass = 0**	**AI Pass = 1**	**Total**
0	23%	9%	32%
1	37%	31%	68%
Total	60%	40%	100%

Alignment in control condition: 54% (When AI fails: 38%; When AI passes: 78%)

Table 2c: T1 - AI Pass vs. Screener Pass in Black box AI Treatment Condition 1

Screener Pass	**AI Pass = 0**	**AI Pass = 1**	**Total**
0	33%	5%	38%
1	22%	40%	62%
Total	55%	45%	100%

Alignment in treatment 1 condition: 73% (When AI fails: 60%; When AI passes: 88%)

Table 2d: T2 - AI Pass vs. Screener Pass in Narrative AI Treatment Condition

Screener Pass	**AI Pass = 0**	**AI Pass = 1**	**Total**
0	38%	5%	43%
1	21%	36%	57%
Total	59%	41%	100%

Alignment in treatment 2 condition: 75% (When AI fails: 65%; When AI passes: 88%)

Table 3. Screener Decisions: LPMs of How AI Narratives Influence Passing Rates

	Dependent variable:					
	Screener Decision (1 = Pass)					
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline = Human-Only Control (C)						
AI Treatment (T1 & T2)	-0.090***					
	(0.018)					
Black box AI (BBAI) T1		-0.065***	-0.076***	-0.068***	-0.052*	-0.071***
		(0.021)	(0.019)	(0.019)	(0.029)	(0.020)
Narrative AI (NAI) T2		-0.115***	-0.117***	-0.115***	-0.124***	-0.101***
		(0.021)	(0.019)	(0.019)	(0.029)	(0.021)
AI Objective Fail			-0.575***	-0.505***	-0.504***	-0.451***
			(0.020)	(0.069)	(0.069)	(0.067)
AI Subjective Fail			-0.255***	-0.385***	-0.384***	-0.313***
			(0.018)	(0.084)	(0.085)	(0.083)
Non-expert			-0.024	-0.020	-0.015	-0.220
			(0.016)	(0.015)	(0.026)	(0.145)
Confidence			0.015***	0.012***	0.012***	0.010**
			(0.005)	(0.005)	(0.005)	(0.005)
Sequence			-0.053***	-0.047***	-0.047***	-0.047***
			(0.016)	(0.015)	(0.015)	(0.015)
BBAI T1 x Non-expert					-0.028	
					(0.038)	
NAI T2 x Non-expert					0.015	
					(0.038)	
Constant	0.683***	0.683***	0.904***	0.942***	0.938***	0.956***
	(0.015)	(0.015)	(0.026)	(0.058)	(0.059)	(0.092)
Observations	3,002	3,002	3,002	3,002	3,002	3,002
Solution Dummies	N	N	N	Y	Y	Y
Screener Dummies	N	N	N	N	N	Y
R^2	0.008	0.010	0.236	0.279	0.279	0.398

*p<0.1; **p<0.05; ***p<0.01

Table 4a. Screener Decisions: LPMs of How AI Narratives Influence Passing Rates by AI Recommendation Type (AI Passes the Solution)

	Dependent variable:				
	Screen Decision (1 = Pass)				
	(1)	(2)	(3)	(4)	(5)
Baseline = Human-only Control (C)					
Black box AI (BBAI) T1	0.104***	0.106***	0.110***	0.074**	0.113***
	(0.024)	(0.024)	(0.024)	(0.038)	(0.028)
Narrative AI (NAI) T2	0.104***	0.103***	0.110***	0.029	0.142***
	(0.025)	(0.025)	(0.025)	(0.039)	(0.029)
Non-expert		-0.023	-0.090***	-0.086**	
		(0.020)	(0.035)	(0.035)	
Confidence		0.001	0.002	0.001	0.003
		(0.006)	(0.006)	(0.006)	(0.007)
Sequence		-0.026	-0.023	-0.023	-0.022
		(0.020)	(0.020)	(0.020)	(0.020)
BBAI T1 x Non-expert				0.061	
				(0.049)	
NAI T2 x Non-expert				0.135***	
				(0.051)	
Constant	0.780***	0.803***	0.867***	0.893***	0.515***
	(0.017)	(0.031)	(0.054)	(0.055)	(0.161)
Observations	1,264	1,264	1,264	1,264	1,264
Solution FE	N	N	Y	Y	Y
Screener FE	N	N	N	N	Y
R^2	0.019	0.021	0.052	0.058	0.308

*p<0.1; **p<0.05; ***p<0.01

Table 4b. Screener Decisions: LPMs of How AI Narratives Influence Passing Rates by AI Recommendation Type (AI Fails the Solution on Objective Criteria)

	Dependent variable:				
	Screen Decision (1 = Pass)				
	(1)	(2)	(3)	(4)	(5)
Baseline = Human-only Control (C)					
Black box AI (BBAI) T1	-0.229***	-0.220***	-0.214***	-0.191***	-0.104
	(0.040)	(0.039)	(0.038)	(0.061)	(0.075)
Narrative AI (NAI) T2	-0.196***	-0.195***	-0.212***	-0.162***	-0.155**
	(0.039)	(0.038)	(0.038)	(0.058)	(0.070)
Non-expert		0.069**	0.068**	0.110**	0.550*
		(0.032)	(0.032)	(0.053)	(0.310)
Confidence		0.037***	0.029***	0.029***	0.014
		(0.010)	(0.009)	(0.009)	(0.011)
Sequence		-0.052	-0.054*	-0.058*	-0.043
		(0.032)	(0.031)	(0.032)	(0.034)
BBAI T1 x Non-expert				-0.040	-0.084
				(0.078)	(0.097)
NAI T2 x Non-expert				-0.088	-0.032
				(0.077)	(0.093)
Constant	0.407***	0.281***	0.199***	0.178**	-0.004
	(0.027)	(0.046)	(0.068)	(0.071)	(0.160)
Observations	720	720	720	720	720
Solution FE	N	N	Y	Y	Y
Screener FE	N	N	N	N	Y
R^2	0.053	0.083	0.157	0.159	0.524

*p<0.1; **p<0.05; ***p<0.01

Table 4c. Screener Decisions: LPMs of How AI Narratives Influence Passing Rates by AI Recommendation Type (AI Fails the Solution on Subjective Criteria)

	Dependent variable:				
	Screen Decision (1 = Pass)				
	(1)	(2)	(3)	(4)	(5)
Baseline = Human-only Control (C)					
Black box AI (BBAI) T1	-0.203***	-0.210***	-0.186***	-0.123**	-0.197***
	(0.036)	(0.036)	(0.035)	(0.054)	(0.059)
Narrative AI (NAI) T2	-0.318***	-0.322***	-0.312***	-0.266***	-0.312***
	(0.036)	(0.036)	(0.035)	(0.054)	(0.061)
Non-expert		-0.100***	-0.087***	-0.026	-0.358
		(0.030)	(0.029)	(0.049)	(0.280)
Confidence		0.011	0.010	0.010	0.011
		(0.010)	(0.009)	(0.009)	(0.010)
Sequence		-0.081	-0.067**	-0.070**	-0.057*
		(0.029)	(0.029)	(0.029)	(0.029)
BBAI T1 x Non-expert				-0.111	-0.005
				(0.072)	(0.083)
NAI T2 x Non-expert				-0.078	0.012
				(0.071)	(0.083)
Constant	0.763***	0.824***	0.828***	0.796***	0.954***
	(0.024)	(0.030)	(0.078)	(0.081)	(0.152)
Observations	1,018	1,018	1,018	1,018	1,018
Solution FE	N	N	Y	Y	Y
Screener FE	N	N	N	N	Y
R^2	0.075	0.086	0.146	0.148	0.451

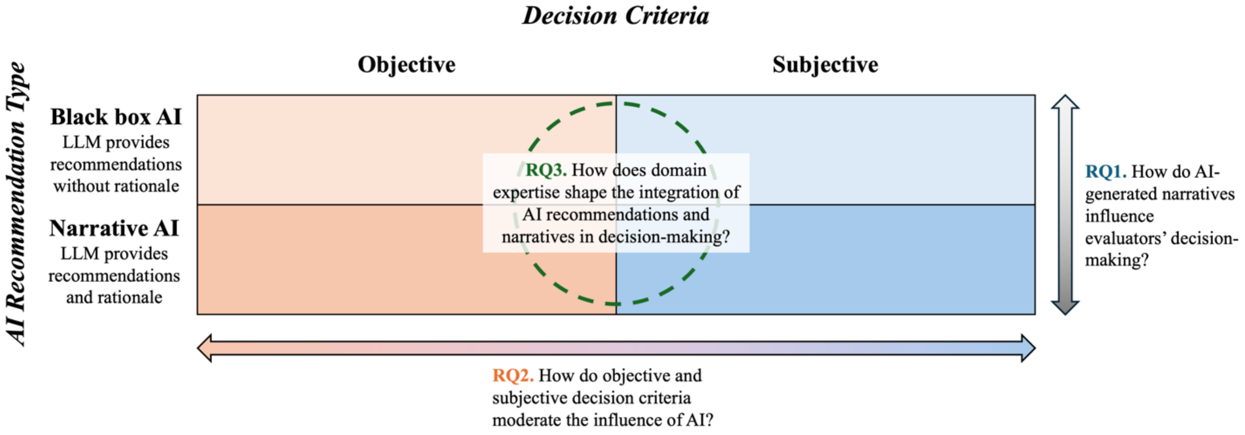
*p<0.1; **p<0.05; ***p<0.01

Table 5. Screener Decisions: LPMs of How AI Narratives Influence Passing Rates by AI Recommendation Type and Screener Mouse Click Behavior

	Dependent variable:				
	Screen Decision (1 = Pass)				
	(1) All	(2) Interaction	(3) AI Pass	(4) AI Objective Fail	(5) AI Subjective Fail
Baseline = Human-only Control (C)					
Black box AI (BBAI) T1	-0.066***	-0.040	-0.070	-0.509**	0.200
	(0.023)	(0.095)	(0.116)	(0.219)	(0.188)
Narrative AI (NAI) T2	-0.111***	-0.099	0.019	-0.774***	-0.060
	(0.023)	(0.096)	(0.117)	(0.224)	(0.188)
Log mouse clicks	-0.203***	-0.197***	-0.210***	-0.394***	-0.100
	(0.021)	(0.035)	(0.043)	(0.084)	(0.067)
Outsider	-0.009	-0.008	-0.015	0.025	-0.038
	(0.022)	(0.022)	(0.028)	(0.047)	(0.043)
Confidence	0.008	0.008	-0.007	0.036***	0.009
	(0.006)	(0.006)	(0.007)	(0.012)	(0.012)
Sequence	-0.075***	-0.075***	-0.050**	-0.066*	-0.098***
	(0.019)	(0.019)	(0.025)	(0.039)	(0.037)
BBAI T1 x Log mouse clicks		-0.015	0.102	0.174	-0.211**
		(0.051)	(0.064)	(0.114)	(0.101)
NAI T2 x Log mouse clicks		-0.007	0.054	0.285**	-0.124
		(0.052)	(0.065)	(0.119)	(0.100)
Constant	1.339***	1.327***	1.262***	0.995***	0.903***
	(0.078)	(0.093)	(0.099)	(0.183)	(0.164)
Observations	1,959	1,959	840	469	650
Solution FE	Y	Y	Y	Y	Y
Evaluator FE	N	N	N	N	N
R^2	0.299	0.299	0.114	0.220	0.175

*p<0.1; **p<0.05; ***p<0.01

Figure 1. A Framework for Human Decision-Making with Objective and Subjective Criteria and LLM Recommendation Types

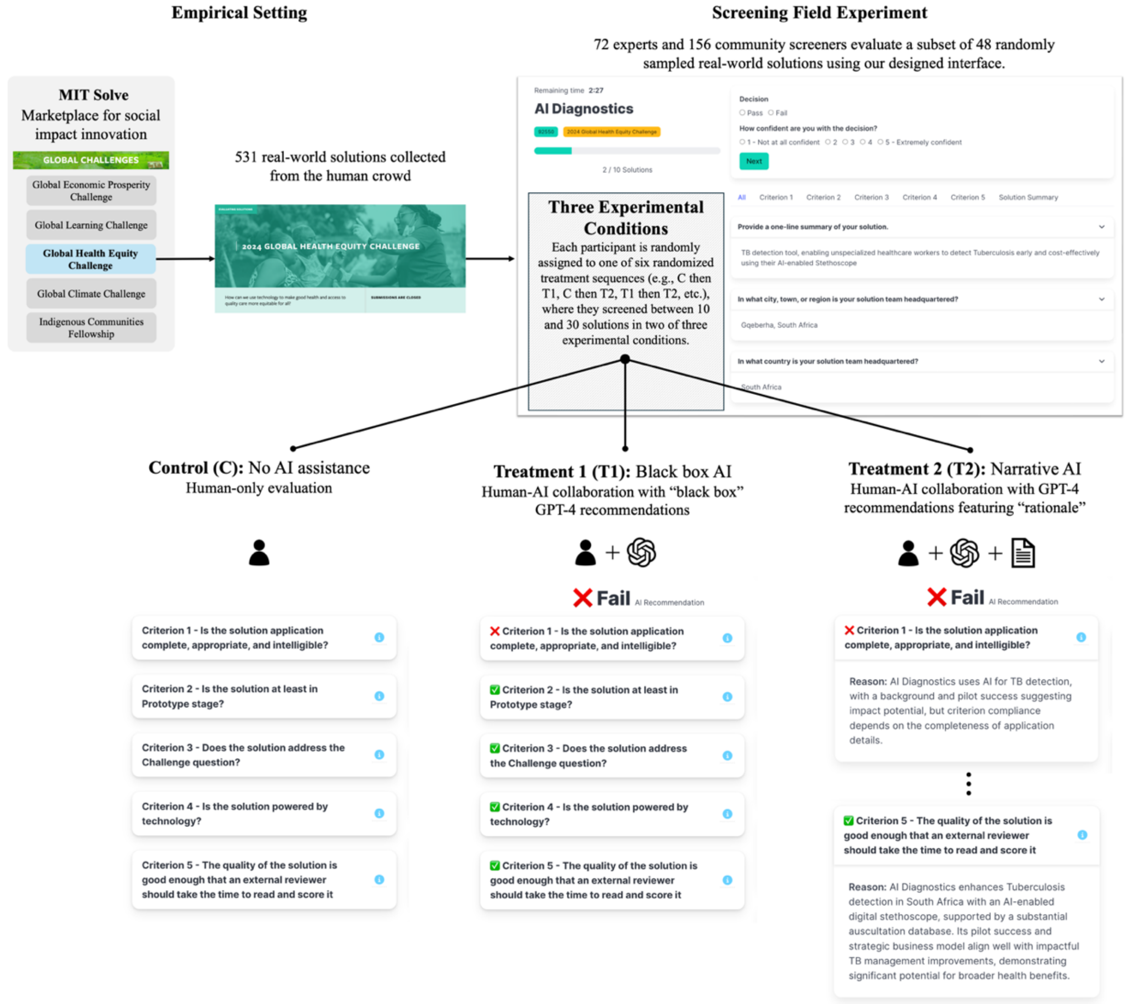

Decision Criteria
Objective
Subjective
AI Recommendation Type
Black box AI
LLM provides recommendations without rationale
Narrative AI
LLM provides recommendations and rationale

RQ3. How does domain expertise shape the integration of AI recommendations and narratives in decision-making?

RQ1. How do AI-generated narratives influence evaluators' decision-making?

RQ2. How do objective and subjective decision criteria moderate the influence of AI?

Figure 2. A Process Flow of the Experimental Design and Web Interface for Solution Screening


Empirical Setting
Screening Field Experiment
MIT Solve
Marketplace for social impact innovation
GLOBAL CHALLENGES
Global Economic Prosperity Challenge
Global Learning Challenge
Global Health Equity Challenge
Global Climate Challenge
Indigenous Communities Fellowship
2024 GLOBAL HEALTH EQUITY CHALLENGE
SUBMISSIONS ARE CLOSED
Remaining time 2:27
AI Diagnostics
92550
2024 Global Health Equity Challenge
2 / 10 Solutions
Decision
○ Pass ○ Fail
○ 1 - Not at all confident ○ 2 ○ 3 ○ 4 ○ 5 - Extremely confident
Next
All
Criterion 1
Criterion 2
Criterion 3
Criterion 4
Criterion 5
Solution Summary
Three Experimental Conditions
Control (C): No AI assistance
Treatment 1 (T1): Black box AI
Treatment 2 (T2): Narrative AI
✗ Fail AI Recommendation
✗ Fail AI Recommendation

531 real-world solutions collected from the human crowd

How can we use technology to make good health and access to quality care more equitable for all?

72 experts and 156 community screeners evaluate a subset of 48 randomly sampled real-world solutions using our designed interface.

How confident are you with the decision?

Provide a one-line summary of your solution.

TB detection tool, enabling unspecialized healthcare workers to detect Tuberculosis early and cost-effectively using their AI-enabled Stethoscope

In what city, town, or region is your solution team headquartered?

Gqeberha, South Africa

In what country is your solution team headquartered?

South Africa

Each participant is randomly assigned to one of six randomized treatment sequences (e.g., C then T1, C then T2, T1 then T2, etc.), where they screened between 10 and 30 solutions in two of three experimental conditions.

Control (C): No AI assistance
Human-only evaluation

- Criterion 1 - Is the solution application complete, appropriate, and intelligible?
- Criterion 2 - Is the solution at least in Prototype stage?
- Criterion 3 - Does the solution address the Challenge question?
- Criterion 4 - Is the solution powered by technology?
- Criterion 5 - The quality of the solution is good enough that an external reviewer should take the time to read and score it

Treatment 1 (T1): Black box AI
Human-AI collaboration with “black box” GPT-4 recommendations

- ✗ Criterion 1 - Is the solution application complete, appropriate, and intelligible?
- ✓ Criterion 2 - Is the solution at least in Prototype stage?
- ✓ Criterion 3 - Does the solution address the Challenge question?
- ✓ Criterion 4 - Is the solution powered by technology?
- ✓ Criterion 5 - The quality of the solution is good enough that an external reviewer should take the time to read and score it

Treatment 2 (T2): Narrative AI
Human-AI collaboration with GPT-4 recommendations featuring “rationale”

- ✗ Criterion 1 - Is the solution application complete, appropriate, and intelligible?

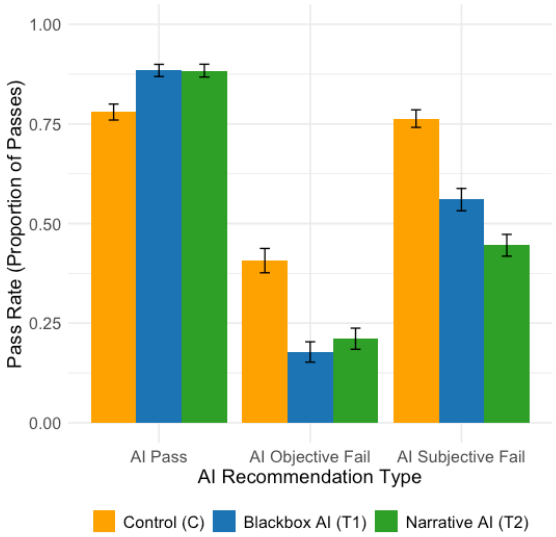
Reason: AI Diagnostics uses AI for TB detection, with a background and pilot success suggesting impact potential, but criterion compliance depends on the completeness of application details.

⋮

- ✓ Criterion 5 - The quality of the solution is good enough that an external reviewer should take the time to read and score it

Reason: AI Diagnostics enhances Tuberculosis detection in South Africa with an AI-enabled digital stethoscope, supported by a substantial auscultation database. Its pilot success and strategic business model align well with impactful TB management improvements, demonstrating significant potential for broader health benefits.

Figure 3. Pass Rate (Proportion of Screening Decision = Pass) by Experimental Condition and AI Recommendation Type



Appendix A

Table A1. Screener Decision Time: OLS Models of How AI Narratives Influence Screener Decision Times by AI Recommendation Type

	Dependent variable:			
	Decision Time (seconds)			
	(1) All	(2) AI Pass	(3) AI Objective Fail	(4) AI Subjective Fail
Black box AI (BBAI) T1	-0.015	-3.802	-3.190	7.793
	(4.031)	(5.888)	(8.181)	(7.396)
Narrative AI (NAI) T2	-1.987	-6.097	-7.449	5.780
	(4.080)	(6.082)	(8.124)	(7.384)
Non-expert	-12.665***	-9.112*	-21.231***	-11.676*
	(3.342)	(4.906)	(6.723)	(6.138)
Confidence	-0.431	-1.763	4.557**	-2.308
	(1.024)	(1.466)	(2.016)	(1.974)
Sequence	-24.527***	-24.829***	-22.236***	-25.810***
	(3.328)	(4.885)	(6.703)	(6.093)
Constant	130.852***	135.683***	145.281***	140.736***
	(12.407)	(12.973)	(14.540)	(16.255)
Observations	2,941	1,238	698	1,005
Solution FE	Y	Y	Y	Y
Evaluator FE	N	N	N	N
R^2	0.046	0.046	0.066	0.035

Note: Table A1 excludes screeners with missing or inaccurate decision times. These fell into two groups: some students refreshed the page or used the back button on their browser, which caused errors in calculating the screening start time, while others left their webpage during the session, resulting in excessively long screening times per solution. *p<0.1; **p<0.05; ***p<0.01

Table A2: Screener Decisions: LPMs of How AI Narratives Influence Passing Rates by AI Recommendation Type and Screener Trust in AI

	Dependent variable:				
	Screen Decision (1 = Pass)				
	(1)	(2)	(3)	(4)	(5)
	Trust	Interaction	AI Pass	AI Objective Fail	AI Subjective Fail
Black box AI (BBAI) T1	-0.067***	-0.021	-0.071	0.071	-0.006
	(0.019)	(0.073)	(0.094)	(0.155)	(0.141)
Narrative AI (NAI) T2	-0.114***	-0.216***	-0.121	-0.277*	-0.270*
	(0.020)	(0.075)	(0.099)	(0.154)	(0.139)
Trust in AI	0.012*	0.009	-0.009	0.006	0.024
	(0.007)	(0.011)	(0.015)	(0.023)	(0.020)
Non-expert	-0.039**	-0.041**	-0.041*	0.075**	-0.122***
	(0.017)	(0.017)	(0.022)	(0.035)	(0.033)
Confidence	0.011**	0.012**	0.001	0.027***	0.011
	(0.005)	(0.005)	(0.006)	(0.010)	(0.010)
Sequence	-0.043***	-0.044***	-0.021	-0.064**	-0.059*

	(0.016)	(0.016)	(0.021)	(0.033)	(0.030)
BBAI T1 x Trust in AI		-0.010	0.043**	-0.067**	-0.042
		(0.016)	(0.020)	(0.033)	(0.031)
NAI T2 x Trust in AI		0.023	0.052**	0.012	-0.008
		(0.016)	(0.021)	(0.034)	(0.030)
Constant	0.917***	0.928***	0.925***	0.175	0.758***
	(0.065)	(0.075)	(0.083)	(0.122)	(0.119)
Observations	2,816	2,816	1,195	678	943
Solution FE	Y	Y	Y	Y	Y
Evaluator FE	N	N	N	N	N
R^2	0.273	0.275	0.066	0.160	0.152

*p<0.1; **p<0.05; ***p<0.01

Table A3. External Evaluations: OLS regression Models of How AI Recommendations Impact the Quality of Screener Decisions

	Dependent variable:			
	Alignment	Innovativeness	Feasibility	Impact
	(1)	(2)	(3)	(4)
Baseline = Human-Only Control				
Black box AI (BBAI) Treatment	-0.043	-0.061	-0.081*	-0.049
	(0.045)	(0.044)	(0.041)	(0.041)
Narrative AI (NAI) Treatment	-0.008	-0.027	-0.100**	-0.011
	(0.044)	(0.043)	(0.040)	(0.040)
Screen Decision = Pass	0.145***	0.153***	0.261***	0.206***
	(0.040)	(0.039)	(0.036)	(0.036)
Non-expert	0.015	0.023	0.020	0.030
	(0.022)	(0.022)	(0.020)	(0.020)
Confidence	0.004	0.003	0.007	0.005
	(0.007)	(0.007)	(0.006)	(0.006)
Sequence	0.042*	0.008	0.026	0.032
	(0.022)	(0.021)	(0.020)	(0.020)
Black box AI x Screen Decision = Pass	0.142**	0.218***	0.162***	0.149***
	(0.056)	(0.054)	(0.051)	(0.051)
NAI x Screen Decision = Pass	0.123**	0.153***	0.183***	0.084*
	(0.055)	(0.054)	(0.051)	(0.051)
Constant	3.893***	3.569***	3.326***	3.641***
	(0.043)	(0.042)	(0.039)	(0.039)
Observations	3,002	3,002	3,002	3,002
Solution Dummies	N	N	N	N
R^2	0.036	0.055	0.102	0.061

Note: The evaluation scores for the alignment, innovativeness, feasibility, and impact of each solution were collected in a separate data collection effort with MIT Solve judges, who were randomized six of the 48 solutions to evaluate.

The scores on each dimension were computed as the average rating across all judges assigned the solution. The interaction term between the experimental condition and the screener decision indicates that the treatment conditions were significantly more likely to pass solutions that were rated as being higher in quality by the MIT Solve judges. *p<0.1; **p<0.05; ***p<0.01

Figure A1. Distribution of PostHog Screener Engagement Behaviors

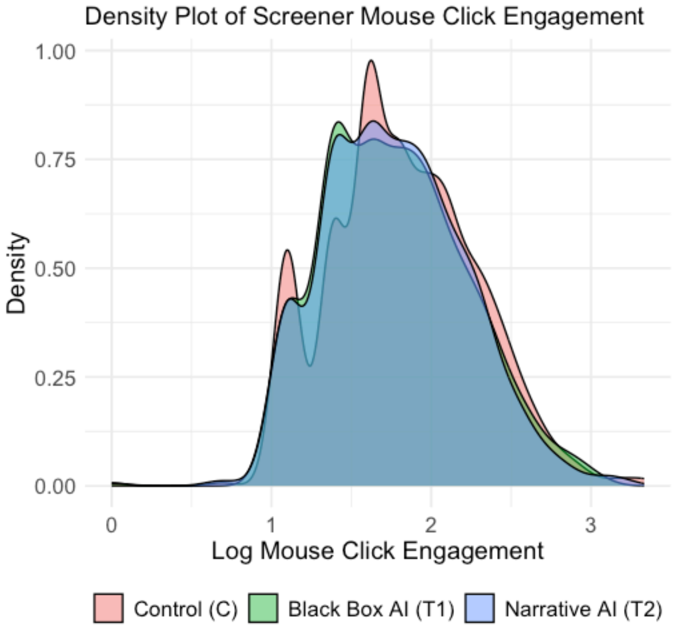
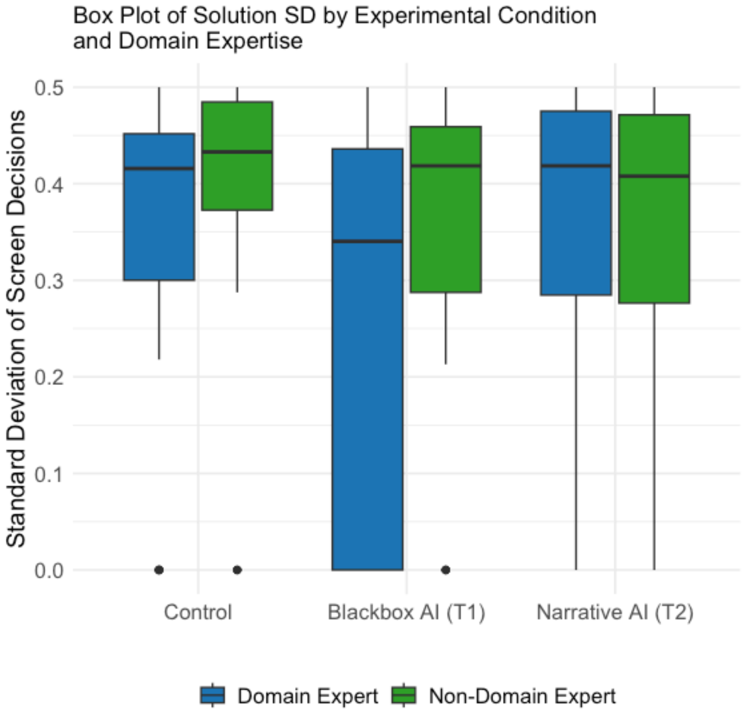


Figure A2. Box Plot of Solution Standard Deviations By Experimental Condition and Domain Expertise



Appendix B

Table B1: Screens per Submission

Group	Mean	Std. Dev	Min	25th Percentile	Median	75th Percentile	Max
Control	23.57	6.09	10	18	21	26	38
Treatment 1	21.38	4.14	10	18	22	23	30
Treatment 2	21.37	4.97	9	16	21	23.25	32
Total	62.48	7.75	40	58	62	67.5	90

Table B2a. Balance tests over solution observables per condition

Variable	Group 1	Group 2	Mean 1	Mean 2	Mean diff
Submitter is a woman	T1	T2	0.432	0.459	-0.027
Submitter is a woman	C	T2	0.414	0.459	-0.045**
Submitter is a woman	C	T1	0.414	0.432	-0.017
Submitter is a US Resident	T1	T2	0.268	0.274	-0.006
Submitter is a US Resident	C	T2	0.24	0.274	-0.034*
Submitter is a US Resident	C	T1	0.24	0.268	-0.028
Submitter had MIT Solve Training course	T1	T2	0.168	0.187	-0.019
Submitter had MIT Solve Training course	C	T2	0.181	0.187	-0.006
Submitter had MIT Solve Training course	C	T1	0.181	0.168	0.013
Submitter is a Past Participant	T1	T2	0.388	0.39	-0.002
Submitter is a Past Participant	C	T2	0.39	0.39	-0.001
Submitter is a Past Participant	C	T1	0.39	0.388	0.002

Table B2b. Balance tests over screener observables per condition

Variable	Group 1	Group 2	Mean 1	Mean 2	Mean diff
Domain expert	T1	T2	1.576	1.569	0.007
Domain expert	C	T2	1.585	1.569	0.016
Domain expert	C	T1	1.585	1.576	0.008
Degree of Use of AI for Personal Use	T1	T2	3.48	3.421	0.058
Degree of Use of AI for Personal Use	C	T2	3.544	3.421	0.122**
Degree of Use of AI for Personal Use	C	T1	3.544	3.48	0.064
Degree of Use of AI for Professional Use	T1	T2	3.655	3.605	0.05
Degree of Use of AI for Professional Use	C	T2	3.627	3.605	0.023
Degree of Use of AI for Professional Use	C	T1	3.627	3.655	-0.028

Used AI for Decision Making in the Past	T1	T2	2.474	2.376	0.098***
Used AI for Decision Making in the Past	C	T2	2.4	2.376	0.024
Used AI for Decision Making in the Past	C	T1	2.4	2.474	-0.073***
Trust in AI	T1	T2	4.41	4.445	-0.035
Trust in AI	C	T2	4.465	4.445	0.02
Trust in AI	C	T1	4.465	4.41	0.055
AI is Reasonable	T1	T2	4.431	4.59	-0.159**
AI is Reasonable	C	T2	4.602	4.59	0.012
AI is Reasonable	C	T1	4.602	4.431	0.171***
Willingness to Use AI in the Future	T1	T2	5.2	5.151	0.048
Willingness to Use AI in the Future	C	T2	5.218	5.151	0.067
Willingness to Use AI in the Future	C	T1	5.218	5.2	0.018

Table B3: Total - AI Decision vs Screener Decision

	AI Accept	AI Reject Obj.	AI Reject Subj.	Total
Screener Accept	35%	7%	20%	62%
Screener Reject Obj.	3%	9%	4%	16%
Screener Reject Subj.	3%	8%	10%	22%
Total	42%	24%	34%	100%

Alignment: 54%

Table B4: Control - AI Decision vs Screener Decision

	AI Accept	AI Reject Obj.	AI Reject Subj.	Total
Screener Accept	32%	10%	27%	68%
Screener Reject Obj.	5%	8%	4%	16%
Screener Reject Subj.	4%	7%	4%	15%
Total	41%	25%	35%	100%

Alignment: 44%

Table B5: T1 - AI Decision vs Screener Decision

	AI Accept	AI Reject Obj.	AI Reject Subj.	Total
Screener Accept	40%	4%	18%	62%
Screener Reject Obj.	2%	9%	4%	14%
Screener Reject Subj.	3%	10%	11%	24%
Total	45%	23%	32%	100%

Alignment: 59%

Table B6: T2 - AI Decision vs Screener Decision

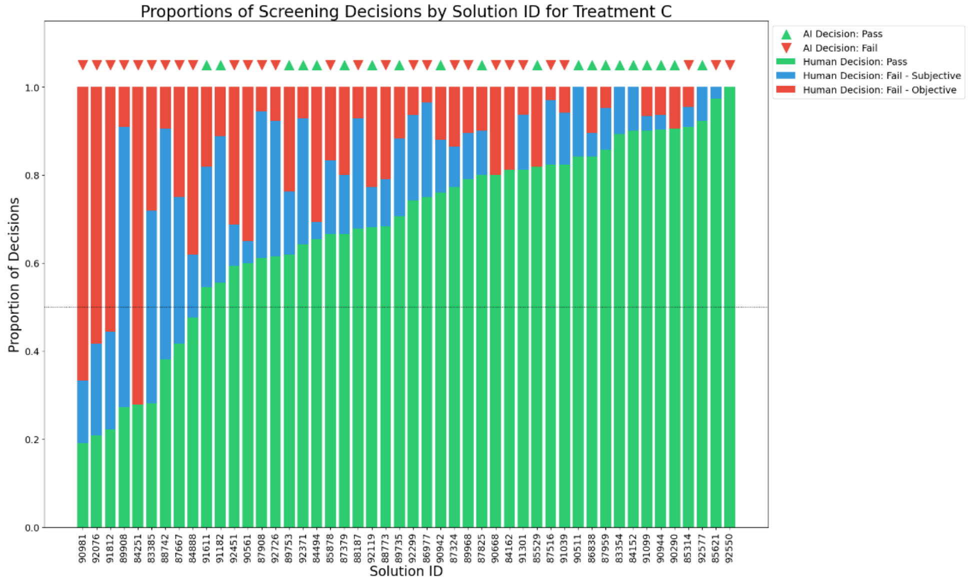
	AI Accept	AI Reject Obj.	AI Reject Subj.	Total
Screener Accept	36%	5%	16%	57%
Screener Reject Obj.	2%	11%	4%	17%
Screener Reject Subj.	3%	9%	15%	26%
Total	41%	25%	35%	100%

Alignment: 61%

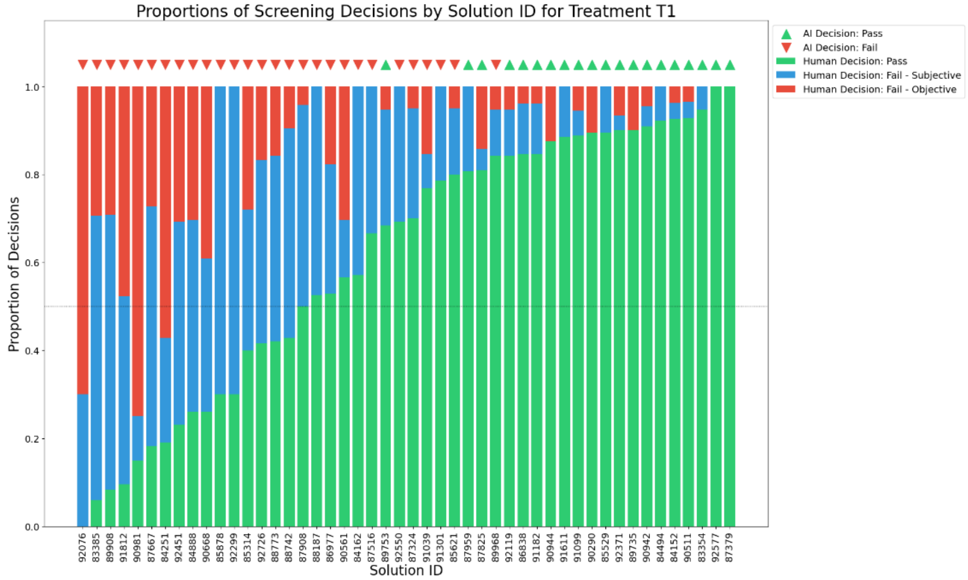
Figure B1: Proportions of Screening Decisions by Solution ID for Each Treatment

This figure compares proportions of screening decisions for the three treatments (C, T1, T2). Stacked bars show the proportion of Pass (green) and Fail outcomes, with Fails subdivided into subjective (blue) and objective (red) criteria. Triangular markers indicate AI decisions (green for Pass, red for Fail). Solutions are ordered by increasing Pass ratio. The 0.5 reference line facilitates cross-treatment comparisons of decision distributions and AI-human alignment.

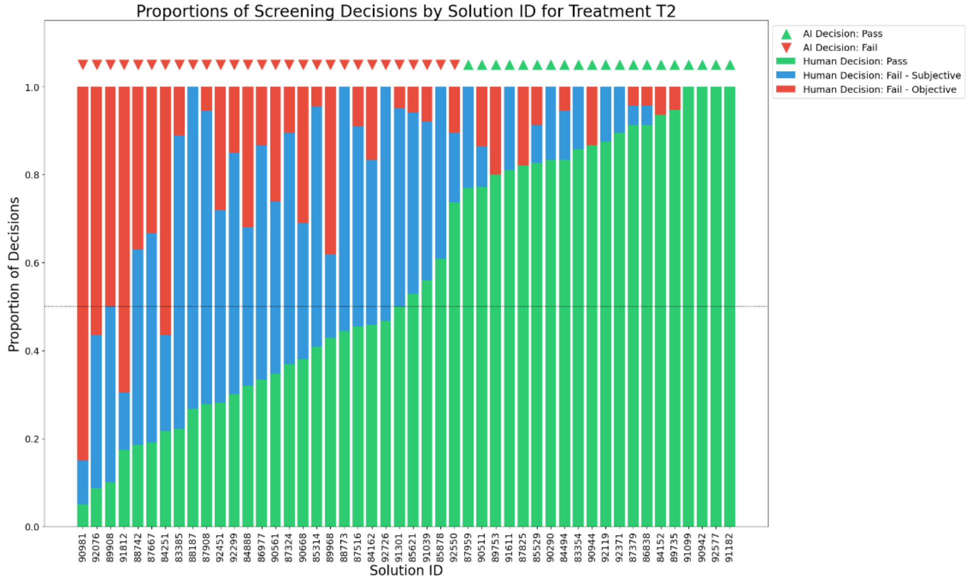
Proportions of Screening Decisions by Solution ID for Treatment C



Proportions of Screening Decisions by Solution ID for Treatment T1



Proportions of Screening Decisions by Solution ID for Treatment T2



Appendix C: General Prompt for Screening

Below, we provide our prompt for GPT-4 using the Python API. We used the `gpt-4-turbo-preview` model version of April 29, 2024, the `client.chat.completions` function calling, and set a temperature of 0.

System prompt:

"You are a screener of startup solutions to the challenge described below. You will screen the solution I give you (at the very end of this text) based on a criterion I give you in detail below. You aim to make sure you let no solution not clearly meeting the criterion pass.

I give you a couple of examples below, one that passes the criterion and one that fails.

Evaluate the solution I give you in the end.

First explain your reasoning step-by-step, and then give your final answer in the template:

Reasoning behind decision: <>

Answer: <probability of being a yes in %>

Confidence in judgment: <>

Put these three elements (reasoning, answer, confidence) in a four-column table with the solution ID for the solution I asked you to evaluate (in column 1) and share that with me.

Make sure you really assess the criterion well, the goal is to screen out and flag the solutions that do not fit the criterion. In column 2, explain the step by step reasoning you follow."

User prompt (adjusted for each criterion to assess)

We bold the part that changes depending on the criterion to assess. The rest remains the same. **> Criterion to assess (adjusted for each prompt):**

"Details of the criterion to consider (see the five possibilities at the end of the prompt)

> Challenge description: "Every person has the right to access the full range of quality health services they need, when and where they're needed. While there has been some progress towards these goals over the last few decades, much of that progress has now slowed or reversed. Currently, half the world lacks access to comprehensive health services. Two billion people face financial hardship due to out-of-pocket healthcare costs. Under-resourced communities (including but not limited to women and girls, ethnic minorities, people with disabilities, and older adults) are often disproportionately affected and experience systematically worse health outcomes.

Technology and innovation have an important role to play in improving health and well-being for all. New technologies can improve health outcomes and access, but only when utilized effectively. Innovation can help these technologies be more affordable, scalable, sustainable, and community-focused. Opportunities for positive change exist across many areas of care including primary care, mental health, and infectious diseases.

MIT Solve seeks exceptional solutions—rooted in and driven by communities—that leverage technology to increase access to quality care and good health. While we are excited to select and support innovators across any health area, we have a particular interest in solutions that:

Ensure health-related data is collected ethically and effectively, and that AI and other insights are accurate, targeted, and actionable.

Increase capacity and resilience of health systems, including workforce, supply chains, and other infrastructure.

Increase access to and quality of health services for medically underserved groups around the world (such as refugees and other displaced people, women and children, older adults, and LGBTQ+ individuals).

”

> Example of solution that passes the criterion:

“Details of filtered out solution from the 2023 Health Challenge inserted here”

> Example of Solution that failed at the criterion:

“Details of a filtered in solution from the 2023 Health Challenge here”

> Solution to evaluate:

“Details of the solution to evaluate”

Details of each criterion integrated into the prompt above

> Criterion 1: “Is the solution application complete, appropriate, and intelligible? Answer no if the application is not in English, or if it provides only a few words for required questions, or if is not intelligible (for e.g., if you can’t figure out what the solution is after reading the application), or if the application was clearly created to offend/isn’t taking the Challenge seriously.”

> Criterion 2: “Is the solution at least in Prototype stage?:Prototype stage means that the venture or organization is building and testing its product, service, or business model. Answer no if no concrete product, service, or business model is being built yet.

Focus on these aspects of the solution in your screening:

-What is your solution’s stage of development?

-Please share details about what makes your solution a Prototype rather than a Concept.
-How many people does your solution currently serve?
-In which countries do you currently operate?
-How are you and your team well-positioned to deliver this solution?”

> Criterion 3: “Does the solution address the Challenge question?: Answer no if the solution does not address the broad Challenge question.

Focus on these aspects of the solution in your screening:
-What is your solution?
-What specific problem are you solving?
-Who does your solution serve, and in what ways will the solution impact their lives?
-Which dimension of the Challenge does your solution most closely address?”

> Criterion 4: “Is the solution powered by technology?: Every solution must include technology, whether new / existing or high-tech / low-tech. Answer no if by removing the tech component of this solution the solution would still work.

Focus on these aspects of the solution in your screening:
-What is your solution?
-What makes your solution innovative?
-Describe the core technology that powers your solution.
-Please select the technologies currently used in your solution”

> Criterion 5: “The quality of the solution is good enough that an external reviewer should take the time to read and score it: Answer no if you think it would be a waste of an external Reviewer’s time to evaluate this solution. The solution should be of very high quality. An indication that it’s not worth the Reviewer’s time: after reading the application, you don’t know what the solution is.”“

Appendix D: Survey Questions and Interview Protocol

Survey Questions

(Only for internal experts): How long have you worked at MIT Solve? (0-1 year/ 1-3 years / 3-5 years/ 5+ years)

How frequently have you used generative AI tools (e.g., ChatGPT, Gemini, Bard, Midjourney, etc) for **personal activities** in the past few months?

I do not use generative AI /

I have tried it once or twice /

I use it occasionally (less than once a week) /

I use it regularly (1-3 times a week) /

I use it very frequently (daily or almost daily) /

How frequently have you used generative AI tools (e.g., ChatGPT, Gemini, Bard, Midjourney, etc) for **professional activities** in the past few months?

I do not use generative AI /

I have tried it once or twice /

I use it occasionally (less than once a week) /

I use it regularly (1-3 times a week) /

I use it very frequently (daily or almost daily) /

Have you previously used Generative AI tools for decision-making or evaluations?

Yes /

No

How much did you trust the AI tool's recommendations during the screening process? (Likert 1-7) No Trust / Neither Trust nor Distrust / Completely Trust

Did you find that the AI provided a reasonable rationale for its decisions? (Likert 1-7) Not Reasonable / Neither Reasonable nor Unreasonable / Completely Reasonable

To what extent would you be willing to use an AI tool to assist with screening decisions in the future? (Likert 1-7) Not Willing At All / Neither Willing nor Unwilling / Extremely Willing

Under which of the following conditions did you find the AI's recommendations most useful? Please select **the most important** reason.

When dealing with information overload. /

When needing to speed up decision-making or response time. /

When looking for insights or patterns in complex data. /

When requiring assistance in unfamiliar topics or areas. /

Other

What did you do when you disagreed with AI's recommendation? (Open ended)

Is there anything else you would like to share with us? (Open ended)

Interview Protocol

Overarching objectives

Human Behavior: Need to really understand the differences between Treatment 1 and 2. How is the rationale, which is present only in Treatment 2, impacting trustworthiness? Are people agreeing or disagreeing with the rationale?

Future: How do people feel about labor replacement for screening? Consider differences between screening vs. evaluations.

General Questions

1. *(Conditional on having experienced in control condition)* Can you describe your approach to screening the solutions in the **human only process**?
2. *(Conditional on having experienced in treatment 1 condition)* Can you describe your approach to screening the solutions with the **AI screener without rationale**?
3. *(Conditional on having experienced in treatment 2 condition)* Can you describe your approach to screening the solutions with the **AI screener with rationale**?
4. Which version do you prefer, and why?
5. How did you use the AI screener in your screening decisions?

    - The pass/fail decision
    - The criteria of failure
    - The rationale for failing/passing a solution
    - AI Summary
    - Selected contents for each criterion

6. Feature Utility:
    - Which feature do you think is most helpful in your decision process and why?
    - Which feature do you think is least useful for you? Why? (we may encounter users who don't know the features)
    - What additional information or features would enable you to assess the effectiveness of the solution more thoroughly?

7. UI/UX Usability:
    - Did you encounter any aspects of the user interface that were unclear, confusing, or difficult to navigate?
    - Based on your interactions with the product, what suggestions do you have for improving the user interface, navigation, or overall design to create a more seamless and enjoyable experience?

Human Behavior:

1. [AI Recommendation Accuracy & Trust]: Did you find the AI recommendations accurate and reliable? **(How) Did you cross-check AI recommendations? How long did it take you to trust AI recommendations?**
2. How did the presence of the AI screener with rationale impact your perception of the reliability of the tool's recommendations, compared to just the AI screener (without rationale)?
3. Did you notice any patterns in your agreement or disagreement with the AI screener with rationale?
4. Reflecting on your experience with both versions (AI screener and AI screener with rationale), do you feel that the rationale provided enhanced or detracted from your trust in the AI screener's recommendations? How?
5. Were there any instances where you found the rationale provided by the AI screener to be unclear or insufficient? How did you navigate such situations during the evaluation process?
6. Can you recall a specific instance where the rationale provided by the AI screener influenced your decision-making process significantly? How did it affect your final screening of the MIT Solve submission?

7. [AI Trust - Behavioral Changes]:
 A. Did you find yourself relying more heavily or less on the AI rationale and recommendations over time?
 B. If so, how did this change your approach to reviewing content and making decisions? such as only reading the summary, just AI rationale, or AI recommendation?
8. Overall, how would you rate your satisfaction with the AI screener's performance in both versions (AI screener and AI screener with rationale)? What improvements or changes would you suggest for future iterations?

Future:

1. How do you foresee the role of human evaluators evolving in light of advancements in AI technology for screening tasks?
2. How do you feel about the potential for AI tools like the one used in this experiment to replace human labor in the screening process of MIT Solve solutions?
3. What concerns, if any, do you have about the implications of adopting AI tools for screening purposes at MIT Solve?
4. [If ethical concerns are not mentioned] Are there any ethical considerations that you believe should be taken into account when implementing AI tools for screening purposes? If so, what are they?
5. What do you view as the most significant strengths and weaknesses of using an AI tool for screening purposes?