Human vs. AI in Startup Evaluation: Evidence from an Early-Stage Accelerator

Andrew Funck*

August 31, 2026

*HEC Paris, Economics and Decision Sciences, and ION Management Science Lab, Paris, andrew.funck@hec.edu

I am grateful to Thomas Astebro for his supervision and guidance. I extend my gratitude to Emmanuel Kemmel and Antonin Bergeaud for their helpful advice and comments. Special thanks go to Julian Chitiva, Andrea Contigiani, Arnaud Dyèvre, Gaetano Gaballo, Francesco Giordano, Juanita González-Uribe, Cedric Gutierrez, Yurii Handziuk, Jamie Hentall-Maccuish, Johan Hombert, Vittoria Iannotta, Michael Impink, Frédéric Koessler, Rui Li, Denisa Mindruta, Naja Pape. Christophe Pérignon, Giacomo Rostagno, Alyssa Rusonik, Carlos Serrano, Tristan Tomala and Johanne Trottin for their valuable feedback. I also thank all participants in HEC PhD seminars and in the Open and User Innovation Conference, Consortium on Competitiveness and Cooperation, and AI Plus Management Doctoral Consortium for feedback and helpful discussions. I am grateful to the research assistants Jhanvi Bansal, Shivangi Kapoor, Vyanktesh Nagre, Vanshika Patel, Apoorv Puranik, who provided valuable support with data collection, and to the managers of the HEC Incubateur, who shared evaluation data used in this research. Any errors are my own. This research was supported by the ION Management Science Lab – HEC Paris and has benefited from a State grant managed by the Agence Nationale de la Recherche under the Investissements d'Avenir programme with the reference ANR-18-EURE-0005 / EUR DATA EFM.

Abstract

Organizations that screen early-stage ventures increasingly draw on artificial intelligence to inform or substitute for expert human judgment, yet little is known about why algorithmic and human assessments diverge. I develop a Bayesian-learning framework in which an evaluator's assessment combines a prior belief about venture quality with a signal extracted from available evidence, weighted by the evaluator's relative confidence in each. The framework implies that if artificial intelligence processes information more precisely, relative to its prior knowledge, than humans, its evaluations should respond more to venture-specific evidence and rely less on priors. I test this hypothesis using 579 applications to a French startup accelerator, comparing more than 24,000 scores assigned by expert human judges on 15 business criteria to counterfactual scores generated for the same applications by a large language model prompted to follow the human grading instructions. Consistent with the framework, AI grades are systematically more responsive than human grades to application content, especially for negative information. Linking scores to venture survival, employment, and capital raised, I find that AI scores are at least as strongly associated with performance as human scores. These results are robust to prompting strategy and look-ahead bias but sensitive to model choice: newer, more capable models converge toward human evaluative patterns with no gain in their association with outcomes. These findings suggest that general purpose AI models can already match the quality of human evaluators in screening ventures.

1. INTRODUCTION

Early-stage venture selection is a central, difficult problem in innovation-driven economies. Startups are widely recognized as key engines of job creation, technological progress, and productivity growth: business startups account for about 20% of US gross job creation (Haltiwanger et al., 2013), and venture capital-backed startups account for 41% of total US stock market capitalization and 62% of publicly traded companies' research and development spending (Bailey et al., 2026). Yet their future performance is highly uncertain at the point when support-allocation decisions are made (Scott et al., 2020). Ventures at early stage typically have no operating history and little hard evidence on which to base a precise evaluation (Damodaran, 2009; Packard et al., 2017; Fisher and Neubert, 2023). As a result, organizations that screen ventures, such as incubators, accelerators, venture capital firms, and public funding agencies, must rely on expert judgment to allocate scarce support on the basis of limited, noisy, and largely qualitative information.

A substantial body of research studies how experts evaluate startups: what criteria they use (MacMillan et al., 1985; Gompers et al., 2020), why they use them the way they do (Franke et al., 2008; Huang and Pearce, 2015), and how well the resulting judgments perform. Structured evaluation can be genuinely informative: expert evaluation can forecast which inventions will succeed commercially (Åstebro and Elhedhli, 2006), judge rankings in pitch competitions predict which ventures go on to raise financing and survive (Howell, 2020), and even the dispersion of judges' evaluations predicts subsequent venture success (Gius, 2025). Yet a substantial body of research documents how challenging the task remains: studies of venture capital screening, business plan competitions, and expert panels consistently find weak correlations between ex ante assessments and ex post outcomes (Zacharakis and Meyer, 2000; Fafchamps and Woodruff, 2017; McKenzie and Sansone, 2019; Åstebro et al., 2026).

Recent advances in artificial intelligence have renewed interest in algorithmic approaches to

judgment and evaluation. AI is increasingly proposed as a tool to assist, augment, or substitute for human decision-makers in high-stakes settings such as hiring, investment, and global management (Lopez-Lira and Tang, 2023; Dell'Acqua et al., 2023; Jabarian and Henkel, 2025). Adoption in venture screening specifically is underway: venture capitalists are incorporating AI tools in the their decision making process, with data-driven approaches reshaping which ventures attract capital (Lyonnet and Stern, 2022; Bonelli, 2026; Vismara et al., 2025). A nascent literature studies AI directly as an evaluator of ideas and ventures. AI models can generate and evaluate business strategies at a level comparable to entrepreneurs and investors (Csaszar et al., 2024), and generative AI applied to strategic decision-making shows similar promise as a standalone evaluative tool (Doshi et al., 2025). Field-experimental evidence shows that AI-augmented screening can match human evaluators on selection quality (Lane et al., 2024), but also exhibits a distinct mean-variance profile relative to human crowds: higher expected financial return but less radical novelty (Boussioux et al., 2024; Grumbach et al., 2026).

Most of existing works compare AI and human evaluations and document how much the two differ, but they rarely offer a rationale for why such differences arise in the first place. In this paper, I propose a conceptual framework for why AI and human evaluators can diverge when assessing a venture, identifying the mechanisms that can lead them to use available evidence differently. I then provide empirical evidence in support of the framework, showing that AI and human evaluators indeed respond differently to the same information about a venture, and I document how their evaluations are, in turn, associated with its future business performance.

I model the venture-evaluation problem as Bayesian learning, in the tradition of DeGroot (2004). The evaluator, human or AI, evaluates a venture on all characteristics informative of commercial success, such as its team, its business model, and its financial prospects. The evaluator starts with a prior belief about each of these venture characteristics: a default expectation, built from its own past knowledge, of what a strong venture looks like on that characteristic, formed

before accessing any information specific to this venture. When accessing venture specific information, the evaluator extracts a signal of the venture's true level on each characteristic, whose precision depends on how costly information processing is for that evaluator. The evaluator's assessment on each dimension is then a weighted average of its prior belief and the signal extracted, with more weight going to whichever of the two it holds with more confidence. This framework gives two distinct reasons why two evaluators looking at the same venture can reach different assessments: (i) they may hold different prior beliefs, or (ii) they may extract signal of different precision from the same information. I hypothesize that AI evaluators have an advantage over humans specifically in processing information, so that the AI should extract a more precise signal; conversely, I expect AI to hold a less precise prior. I therefore expect AI scores to depend more heavily on the specific information available about a venture, and human scores to depend comparatively more on their prior beliefs.

I test this conceptual framework in the context of admissions to an accelerator program funded by a leading business school in France. Accelerators are a particularly relevant setting for studying ventures evaluation. Beyond offering mentoring, networks, and resources, they function as centralized selection devices, screening large numbers of early-stage ventures on behalf of the investors who rely on their certification (Cohen, 2014; Cohen et al., 2019). Their decisions carry real stakes: participation can improve venture performance (Hallen et al., 2020; Gonzalez-Uribe and Leatherbee, 2018; Yu, 2020), and screening quality is itself informative, since judge scores in accelerator and competition settings predict which ventures go on to become high-growth firms (Gonzalez-Uribe and Reyes, 2021).

In the admission process of this accelerator, human judges score text applications of ventures against a fixed set of evaluation criteria to determine program admission. Using the same applications, the same criteria, and the same discrete scoring scale, I generate counterfactual evaluations of each application by prompting a state-of-the-art Large Language Model (LLM) to follow the

human grading instructions, then compare the two sets of scores directly. In my first set of empirical results, drawing on 33,000 scores assigned by human judges and the AI across 15 criteria for 579 ventures, I document that the AI scores are more responsive than human judges to application information. Using a text-based measure of how strongly an application reads as positive or negative on each criterion, I find that both evaluators respond to it, but the AI's response exceeds the human benchmark by 34%, driven in particular by a stronger response to negative information.

To assess how AI and human evaluations relate to venture performance, I manually collected post-evaluation outcomes, survival status, total employment, and total capital raised, for the entire sample of ventures from publicly available sources, and estimated the relationship between grades and these outcomes. In my second set of empirical results, I show that AI scores are at least as strongly associated with venture performance as human scores, particularly for employment. This association holds across the outcome distribution, not just on average; is strongest at the top of the distribution; holds whether considering the aggregate of evaluations for a venture or criterion by criterion, and whether estimated using OLS or rank correlations; is robust to measuring venture performance at two separate time horizons; and is not an artifact of look-ahead bias.

Following (Carlson and Burbano, 2026) on the importance of testing the sensitivity of LLM-based results to configuration choices, I check robustness of my empirical results across prompting strategies and AI models. Prompting does not matter, model choice does: more recent models evaluate ventures more similarly to human judges,with AI's larger response to available information narrowing, and no corresponding improvement in the association with venture performance.

These results carry a direct implication for how screening organizations should allocate human effort. If AI evaluations are associated with venture performance at least as strongly as expert human evaluations, then human judgment may not be needed for ranking ventures. The framework I propose points towards a wider interpretation: when AI and human assessments diverge system-

atically because of different priors and different precision in processing available evidence, each carries valuable information the other does not, and the design problem should become identifying when combining them outperforms using either alone (Boussioux et al., 2024; Dell'Acqua et al., 2025) and when it does not (Vaccaro et al., 2024). This problem is not specific to accelerators: the same considerations apply to many other contexts where evaluators must judge on limited and largely qualitative evidence under substantial uncertainty about eventual outcomes, as in hiring or early-stage investment.

This paper contributes to three streams of literature. First, I contribute to the literature on venture evaluation (Baum and Silverman, 2004; Song et al., 2008), and in particular to the tradition of Bayesian entrepreneurship, which models entrepreneurial and evaluation decisions as a process of belief updating under uncertainty (Camuffo et al., 2020; Åstebro et al., 2025; Agrawal et al., 2025). I extend this tradition to the case of AI as an evaluator, contributing to the broader literature on AI decision-making (Kleinberg et al., 2018; Agrawal et al., 2018; Agarwal et al., 2025; Bryan and Gans, 2026). Second, the paper speaks to the empirical literature on algorithmic strategic evaluation (Shepherd and Majchrzak, 2022; Cao et al., 2024; Kleinert and Urbig, 2026; Li et al., 2026; Csaszar et al., 2026). I document where AI and human evaluations of the same ventures diverge, and assess the quality of both by associating them with several measures of venture performance for a large sample of companies spanning different sectors, maturities, and business models. I further show how this divergence, and the quality of the AI evaluations vary across generations of AI models. Third, I contribute to the literature on accelerators (Wright, 2018; Hallen et al., 2020; Gonzalez-Uribe and Reyes, 2021), showing how human evaluators use the information available to them when selecting ventures in the context of an early stage accelerator(Åstebro et al., 2025, 2026), to what extent their evaluations relate to venture performance, and that automation of the selection process would not decrease the quality of selected ventures.

The rest of the paper is structured as follows: Section 2 presents the conceptual framework.

Section 3 describes the institutional context and data. Section 4 presents the empirical results. Section 5 provides robustness tests. Section 6 concludes with a discussion of limitations, avenues for future research, and managerial implications.

2. CONCEPTUAL FRAMEWORK

I develop a conceptual framework for how human experts and AI models evaluate early-stage ventures. I begin by describing the evaluation problem faced by any evaluator of early-stage ventures, human or AI, arguing that human and AI evaluators are expected to diverge along two specific structural dimensions. I then formalize this argument through a stylized Bayesian-learning model. Based on this discussion, I derive the model's central hypothesis.

2.1. The Evaluation Problem

A substantial empirical literature on new venture performance has long identified a set of observable features most informative of which ventures go on to survive and grow (Baum and Silverman, 2004; Song et al., 2008). These typically include features such as the strength of a venture's team, the viability of its business model, the size of its addressable market, the strength of its competitive advantage, the maturity of the technical development of is product, and the venture's financing prospects (MacMillan et al., 1985) At the earliest stages, when no business outcome has yet been realized, evaluating a venture, whether the evaluator is a human expert or an AI model, amounts then to forming a judgment, from the information available, about the level of the venture on each of these characteristics (Franke et al., 2008).

I model this evaluation through the lens of Bayesian learning as an inference problem in which every evaluator combines prior knowledge and available information to form an assessment. I characterize human and AI judges as two types of evaluators within this common framework. Evaluators are thus characterized by two aspects: a prior belief about what a strong venture looks like on each characteristic, held before observing any information about the specific venture at hand, and

an ability to extract venture-specific information from what is available, which determines how far a specific piece of evidence can move the evaluator away from that prior. Evaluators differ only along these two dimensions, and the same framework could equally be used to compare two human judges, or two AI models.

A human evaluator's prior belief about what a strong venture looks like is shaped by direct professional experience, the population of ventures, founders, and business models the evaluator has been exposed to over a career of screening similar proposals (Franke et al., 2008). An AI model's prior is shaped by the statistical regularities encoded during training over a large and heterogeneous corpus of text (Biever, 2023).

The two evaluator types are also expected to differ in their ability to extract venture-specific information. Careful evaluation is not free for a human expert: reading and weighing information carefully is effortful and time-constrained attention is a scarce resource(Simon, 1971; Kahneman, 1973). Evidence from expert judgment in high-stakes evaluative settings shows that the quality of an expert's information processing can degrade over the course of a session as attention is depleted (Kahneman, 2011; Danziger et al., 2011). Experienced evaluators also rely heavily on accumulated pattern recognition, sometimes described as gut feel, precisely to economize on the effort of processing everything in front of them (Huang and Pearce, 2015). An AI model's marginal cost of processing an additional piece of information is, by contrast, close to negligible (Agrawal et al., 2018).

An evaluator's assessment of a venture is therefore the product of these two ingredients: what the evaluator already believed before encountering this particular venture, and what the evaluator was able to extract from evidence about it. I formalize this logic through a stylized Bayesian-learning model, in the tradition surveyed by Baley and Veldkamp (2023), and use it to show how differences between AI and human evaluators along these two dimensions translate into diver-

gence in their assessments of the same venture. The model is intentionally minimal, with standard parametric assumptions chosen to ensure tractability and make the two channels of evaluator heterogeneity sharp and clear. Technical details, derivations, proofs and extensions are collected in Appendix A.

2.2. Model

A venture i is evaluated by two types of evaluators j, human or AI. A venture is characterized by a vector of K latent qualities $\mathbf{q}_i \in \mathbb{R}^K$: the business model, the size of the addressable market, and similar dimensions. Assume withouth loss of generality $K = 2$. Latent quality is drawn from a jointly Gaussian population distribution,

$$\mathbf{q}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_q), \qquad \Sigma_q = \sigma_q^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad \rho \in (-1, 1),$$

Each evaluator assess each venture along each characteristic k choosing a score $m_{ijk} \in \mathbb{R}$ to come as close as possible, in expectation, to the venture's true quality on that characteristic, given whatever information the evaluator has processed:

$$m^*_{ijk} = \arg\min_{m \in \mathbb{R}} \mathbb{E}\left[(m - q_{ik})^2 \mid \mathscr{I}_j(i)\right],$$

where $\mathscr{I}_j(i)$ denotes evaluator j's information set after processing venture i. This quadratic-loss objective is common to both evaluator types, human and AI.

Before observing anything about venture i, every evaluator carries a baseline expectation about each characteristic, formed from everything the evaluator has been exposed to prior to this specific

venture.

$$q_{ik} \sim \mathcal{N}(\mu_{jk}, \tau_{0jk}^{-1}),$$

where the prior mean μ_{jk} encodes the evaluator's default expectation about characteristic k, and the prior precision τ_{0jk} encodes how confidently that default is held.

When an evaluator accesses the information available about a venture, the evaluator extracts a noisy reading s_{ijk} of the venture's latent quality q_{ik} on each characteristic. The noise ε_{ijk} in this reading reflects the limits of the evaluator's own ability to extract the relevant content from what is available.

$$s_{ijk} = q_{ik} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim \mathcal{N}(0, \tau_{\varepsilon j}^{-1}), \qquad k \in \{1,2\}.$$

Signal precision $\tau_{\varepsilon j}$ is common across characteristics and is the second primitive, alongside the prior, that distinguishes one evaluator from another. A high $\tau_{\varepsilon j}$ means the evaluator extracts a precise reading of the available information; a low $\tau_{\varepsilon j}$ means the reading is noisy. $\tau_{\varepsilon j}$ is higher when the evaluator's marginal cost of processing information is low relative to the confidence already held in the prior (formalized in Appendix A.2.2 following Sims, 2003; Matějka and McKay, 2015).

Upon extracting the signal for each characteristic, the evaluator combines it with prior beliefs via Bayes' rule and reports the resulting posterior as the optimal score.

Proposition 1 (Optimal grade)**.** *Under the venture quality and evaluators' objective, prior, and signal structure assumed, the optimal evaluation is*

$$\boxed{m_{ijk}^{*} \;=\; \mathbb{E}[q_{ik} \mid \mathbf{s}_{ij}] \;=\; \underbrace{(1-\omega_{jk})\,\mu_{jk}}_{\textit{prior component}} + \underbrace{\omega_{jk}\, s_{ijk}}_{\textit{signal response}}} \tag{1}$$

with

$$\omega_{jk} \equiv \frac{\tau_{\varepsilon j}}{\tau_{0jk} + \tau_{\varepsilon j}} \in (0,1).$$

The optimal score decomposes additively into two terms. The prior component $(1-\omega_{jk})\mu_{jk}$ reflects the evaluator's default belief about characteristic k, discounted by the relative precision of the prior to the signal. The signal response $\omega_{jk}s_{ijk}$ scales the venture-specific signal by the weight ω_{jk}, which depends on the evaluator's signal precision relative to their prior precision. An evaluator with very high information processing cost extracts almost no information from the available evidence, $\omega_{jk} \to 0$ and reports the same baseline score regardless of the venture; one who extracts information with near-perfect precision, $\omega_{jk} \to 1$, lets the evidence fully determine the score.

2.3. Hypothesis

The model attributes any divergence between AI and human judgment to two sources: differences in priors, and differences in the precision with which information is extracted. Together, these determine how heavily an assessment relies on available evidence relative to the prior. By Proposition 1, the weight ω_{jk} on available evidence is higher whenever an evaluator's signal precision is higher relative to its prior precision, so any force that sharpens an evaluator's reading of a venture, or loosens its attachment to a default expectation, shifts the assessment toward the signal and away from the prior. I expect this ratio, and hence the resulting weight, to be larger for AI than for human evaluators, for reasons attached to each primitive.

On the prior side, a human evaluator's prior is calibrated to the specific task and population of ventures encountered over a career, whereas an AI model's prior is shaped by a broad, generic training corpus. On the signal side, the cost of processing a given piece of information is for an AI model far lower than for a human evaluator, whose attention is limited and prone to bias and fatigue. Overall then, I expect

$$\omega_{AI,k} \equiv \frac{\tau_{\varepsilon,AI}}{\tau_{0,AI,k}+\tau_{\varepsilon,AI}} > \frac{\tau_{\varepsilon,H}}{\tau_{0,H,k}+\tau_{\varepsilon,H}} \equiv \omega_{H,k}$$

giving the model's central hypothesis.

Hypothesis. *The AI evaluator's assessments place greater weight on the signal extracted from available information than the human evaluator's assessments.*

3. CONTEXT AND DATA

This section describes the empirical setting and the data used to test the predictions of the conceptual framework and to assess the quality of AI and human evaluations against venture performance. I first introduce the French accelerator where the evaluations take place. I then describe the sample of ventures evaluated and their performance data. Third, I describe the process used to generate AI evaluations. Last, I present the procedure used to measure the information content of venture applications.

3.1. The HEC Incubateur

The empirical setting of this study is the HEC Incubateur, a Paris-based startup accelerator operated by a leading bussines school in France. The incubator supports early-stage ventures through a four-month acceleration program that combines mentoring, workspace at Station F, one of the world's largest startup campuses, and access to investors and the broader entrepreneurial ecosystem. It is directed primarily to young entrepreneurs living in Paris who have recently graduated from university, often from an entrepreneurship program at HEC, or from its partner institutions. Founded in 2007, the program has supported more than 900 ventures, which collectively have raised over €1 billion in external funding.[1] The incubator runs competitive admission rounds on a quarterly basis (referred to as batches). In each batch, a limited number of ventures are admitted (typically around ten), depending on capacity constraints.

[1] https://builders.hec.fr/incubateur

The admission process unfolds in multiple stages. This paper focuses exclusively on the first-stage evaluation. Entrepreneurs who satisfy basic eligibility criteria are invited to complete a long written application (on average, application files contains 2600 words) consisting of seventy-four questions that cover the principal dimensions of an early-stage venture: team composition, education, and experience; product or technology; customer problem and target market; competitive landscape; business model; development stage; financing resources and needs; and legal status and regulatory constraints.The complete list of application questions is reported for reference in Appendix E.

Each eligible application is independently evaluated by an average of three expert human judges, who are randomly assigned and evaluate applications separately. Judges include former entrepreneurs, investors, academics, and accelerator staff. Judges are not informed about the identity or opinions of the other judges assigned to the same application. In the evaluation form, judges assess ventures on a fixed set of 15 criteria using a discrete 1–3 scoring scale, where 1 denotes low quality, 2 intermediate quality, and 3 high quality. Each score level within a criterion is itself defined in the form by a short textual description of what that level represents. The criteria span core dimensions of venture quality, including team capabilities, clarity of the user pain point, feasibility of the solution, competitive advantage, business model coherence, technical maturity, and financing prospects. The complete list of evaluation criteria and their grading rubric is reported for reference in Appendix F. Within each batch, criterion grades are aggregated to form an overall ranking of ventures. Top-ranked ventures are invited to proceed to the final selection stage.

3.2. Sample of ventures

I study the full population of all 579 applications submitted to the HEC Incubateur and evaluated by human experts across nine batches between Summer 2021 and Winter 2024.

– INSERT TABLE 1 ABOUT HERE –

Table 1 presents descriptive statistics on venture characteristics extracted from applications.

Forty percent of ventures have at least one founder who graduated from HEC Paris, and 46% include at least one female co-founder. Founders are on average 33 years old, and teams typically consist of four members. In terms of organizational maturity, 64% of ventures are already legally incorporated at the time of application. Nearly half (46%) report having developed a minimum viable product, 28% are at the prototype stage, and 26% report a fully working product. Business models are predominantly digital and service-oriented: subscription-based SaaS models account for 54% of ventures, followed by marketplaces (21%) and e-commerce (15%). The remaining ventures operate as agencies or training centers, mobile applications, or media and social platforms. The sample spans a broad range of industries: software and IT is the single largest sector at 24% of ventures, followed by finance and real estate (12%) and life sciences and healthcare (11%), with the remainder distributed across hospitality and education, professional services, retail and distribution, consumer goods, and media and entertainment.

I construct measures of ex-post venture performance using publicly available data collected in Summer 2024. For each venture, I manually gathered information on (i) survival, (ii) employment, measured by the number of full-time equivalent employees (FTEs), and (iii) external funding, measured as the total amount of capital raised since application. Data collection followed a structured coding protocol,[2] under which I searched each venture's name, website, and founders' names across the current and historical web, as well as on LinkedIn, Crunchbase, and, for incorporated French ventures, the INSEE business register, and reported only figures backed by a verifiable source. Outcome data are necessarily incomplete and subject to right censoring, particularly for more recent batches and for funding outcomes, which are relatively rare events at early stages. When no information could be found across sources, ventures are treated as inactive, following standard practice in the literature (Lerner and Malmendier, 2013).

– INSERT TABLE 2 ABOUT HERE –

Table 2 reports outcome summary statistics by batch and for the pooled sample. As of mid-

[2] Available upon request.

2024, approximately 73% of ventures remain active. Among surviving ventures, the average employment level is 6.7 FTEs, with substantial right skewness. About 11% of ventures have raised external funding, and among those that do, the average amount raised is €3.5 million, again with significant dispersion driven by a small number of high-performing outliers.

3.3. AI-Generated Venture Evaluations

To compare human and algorithmic evaluations, I generate counterfactual venture evaluations using a text generation Large Language Model (LLM) that mirrors as closely as possible the information set and evaluation framework faced by human judges. For each venture independently, the AI is provided with the exact same written application submitted to the HEC Incubateur that human judges evaluate. The model is also given the same evaluation schema used by human judges. The baseline prompt casts the model in the role of a reviewer at the HEC Incubateur assessing whether a venture should be admitted to the final stage of the selection process, and restricts its assessment to the content of the application and to information that would have been available as of the application's submission date. For each of the fifteen criteria, the model is instructed to return a discrete score, a confidence distribution over the three possible score levels, and the identifiers of the three application questions whose answers most support the assigned score. The exact text of the prompt is reported in Appendix G. Importantly, the model is not fine-tuned on HEC Incubateur data and is not exposed to any examples of past human scores. All evaluations are generated in a zero-shot setting, directly mapping application text to criterion-level scores.

The baesline model I consider for this task is Llama-3.3-70B-Instruct, a state-of-the-art, publicly available general purpose LLM developped by Meta and pre-trained on a broad mixture of publicly available online text. The model is run with decoding temperature set to zero, which maximizes the reproducibility of its output (Törnberg, 2024). The model's output is not fully deterministic: repeated queries on the same venture return a different score on approximately 2.5% of criterion-venture evaluations. To account for this residual stochasticity, each venture is evaluated

five times, and the final AI score assigned to each venture-criterion pair is the median score across the five runs. Section 5.1 reconsiders this baseline configuration, proposing alternative prompting strategies and models. This procedure yields a score for each of the 15 criteria for each of the 579 ventures, for a total of 8,685 AI scores. Combined with the 24,315 human scores assigned over the same criteria and ventures, the two sets together form a sample of 33,000 evaluations.

3.4. Signal mesure

To test the model's prediction that human and AI grades differ in their response to venture-specific information (Section 2), I construct a measure of the directional information contained in the application of each venture i on each criterion k. The construction exploits the fact that the discrete scores $\{1,2,3\}$ on each criterion are defined verbally in the grading protocol, with each score accompanied by a textual description of what it represents (see Appendix F). For each criterion $\times$ score pair, I represent the verbal score description as a vector using a sentence-embedding model (multilingual-e5-large), distinct from the model used to generate the evaluations. Embedding models map text into a high-dimensional vector space in which position encodes meaning: texts with similar meaning are placed close together, and the direction and distance between two vectors reflect the semantic relationship between the texts they represent (Mikolov et al., 2013). I denote the embeddings of the top- and bottom-score descriptions on each criterion e_k^3 and e_k^1.

I then embed every answer in each venture i's application using the same sentence-embedding model, denoting the embedding of answer q as a_{iq}. For each venture, I locate the single application answer most similar to the top-score description and the single application answer most similar to the bottom-score description, and record the corresponding cosine similarities:

$$s_{ik}^{+} = \max_{q} \operatorname{sim}\left(e_k^3, a_{iq}\right), \qquad s_{ik}^{-} = \max_{q} \operatorname{sim}\left(e_k^1, a_{iq}\right),$$

The quantity s_{ik}^{+} measures how closely the application's most relevant answer matches a high-score

profile on criterion k, and s_{ik}^{-} how closely it matches a low-score profile. The signal is the difference between the two:

$$s_{ik} = s_{ik}^{+} - s_{ik}^{-}.$$

By construction, s_{ik} is positive when the application reads as closer to a strong than to a weak profile on criterion k, negative when the reverse holds, and zero when the two are equally close. The measure thus captures the direction of the criterion-relevant information, that is, whether the application reads as positive or negative on criterion k, rather than the mere presence of content about that criterion.

4. RESULTS

This section presents the paper's central empirical results. I first document the existence and magnitude of systematic differences between AI and human evaluations. I then turn to the source of this divergence, testing the hypothesis derived in Section 2.3 that AI evaluators place more weight than human evaluators on the information extracted from textual descriptions of ventures. Finally, I ask whether the two forms of evaluation differ in their informativeness for venture performance, relating AI and human scores to ex-post outcomes.

4.1. Descriptive Evidence

I begin by comparing the overall grading behavior of the AI to that of human judges. To do so, I estimate two simple decompositions of grade $X_{i,j,k}$, assigned by judge j (human or AI) to venture i on criterion k.

First, I decompose grades into venture and criterion fixed effects, estimated separately for the AI and for the pool of human judges,

$$X_{i,j,k} = \gamma_i + \delta_k + \varepsilon_{ijk}, \qquad j \in \{AI, H\},$$

where γ_i is a venture fixed effect and δ_k is a criterion fixed effect. I then decompose the total sum of squares of $X_{i,j,k}$ into the shares attributable to γ_i, to δ_k, and to the residual ε_{ijk}, separately for each evaluator type. This exercise documents how AI and human evaluations distribute variation across ventures versus criteria. The two evaluator types allocate this variance along opposite dimensions. For human judges, ventures account for 45% of grade variance and criteria for only 15%; for the AI, this pattern inverts, with ventures accounting for 15% of variance and criteria for 45%. Human judges thus differentiate primarily between ventures, grading a given venture relatively uniformly across its criteria, whereas the AI differentiates sharply between criteria within a venture while compressing differences across ventures.

The second decomposition turns to a different question: whether, conditional on the venture and criterion being evaluated, the AI grades at a different level than human judges. I estimate a pooled model with venture $\times$ criterion fixed effects θ_{ik}, which absorb the quality of each venture on each criterion, and evaluator fixed effects α_j, which measure each evaluator's systematic tendency to grade above or below others on the same venture $\times$ criterion cell.

$$X_{i,j,k} = \alpha_j + \theta_{ik} + \varepsilon_{i,j,k}.$$

Because θ_{ik} nets out any genuine cross-venture or cross-criterion differences in quality, α_j isolates leniency specific to the evaluator. Figure 1 reports the distribution of these evaluator fixed effects.

— INSERT FIGURE 1 ABOUT HERE —

The AI's fixed effect lies in the extreme upper tail of the distribution. Holding constant the venture and criterion being evaluated, the AI grades systematically higher than every human judge.

This aggregate gap masks substantial heterogeneity across evaluation dimensions. Figure 2 reports the average relative difference betweeen the AI and the human grade on each of the fifteen criteria (Figure D.1 in Appendix reports differences in levels).

— INSERT FIGURE 2 ABOUT HERE —

The criterion-level differences are large and systematic. Relative to the human mean, the AI–human gap ranges from approximately -15% on financing to more than $+50\%$ on competitive advantage, averaging about $+17\%$: the AI is not uniformly more lenient.

These differences might in principle reflect nothing more than the disagreement routinely observed among human judges, with the AI behaving as one more, somewhat lenient, draw from the human pool. Two further results show that this is not the case. First, I decompose the mean squared error between AI and human grades into the variance of human grades, capturing disagreement among human judges, and the squared distance between the average human grade and the AI grade, capturing systematic bias between the two evaluator types; roughly 75% of the total is attributable to systematic bias rather than to human inter-rater noise (Table C.1 in Appendix C). Second, the correlation between the AI's grades and the average human grade within each venture $\times$ criterion cell is systematically lower than the corresponding correlation for the typical human judge (Figure D.2 in Appendix D): the AI agrees with the human consensus way less than human judges agree with one another.

The results paint a consistent picture: AI grades differ from human grades both on average and criterion by criterion; they vary more across criteria than across ventures; they are higher than the grades of nearly every human judge; and they diverge systematically from the human consensus.

4.2. Information response

Section 2 attributes divergence between human and AI evaluators to two structural sources: differences in prior beliefs about a venture's quality, and differences in the precision with which each evaluator extracts venture-specific information from the evidence available about it. The model's central prediction, in Subsection 2.3, is that the AI's signal precision is larger relative to its prior precision than is the case for human evaluators, because the AI's lower marginal cost of processing

information lets it extract a sharper reading of the venture. A sharper relative signal implies that the AI places more weight on the information it extracts, and correspondingly less on its prior, than human evaluators do. I test this prediction estimating

$$\begin{aligned} X_{ijk} = \alpha + \alpha_{\text{AI}} D_{j=\text{AI}} + \beta_H s_{ik} + \beta_{AI} (D_{j=\text{AI}} \times s_{ik}) \\ + \gamma_H \mathbf{Z_i} + \gamma_{AI} (D_{j=\text{AI}} \times \mathbf{Z_i}) + \delta_{jk} + \lambda_{ij} + \varepsilon_{ijk}, \end{aligned} \tag{2}$$

where s_{ik} measures the signal on venture i about criterion k described in section 3.4; $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation including team size and experience, founder education title, gender and years of experience, number of customer interviewed, capital raised, and incorporation status; δ_{jk} are judge $\times$ criterion fixed effects, and λ_{ij} are judge $\times$ venture fixed effects, included one at a time. The coefficient β_H is the human grade response to the signal and β_{AI} the AI's additional response. The signal is standardized within criterion, so the coefficients measure the change in the grade associated with a one-standard-deviation increase in positive criterion-relevant content relative to the criterion-specific average signal level. With judge $\times$ criterion fixed effects, β_H and β_{AI} are identified from variation in s_{ik} across ventures within a criterion; with judge $\times$ venture fixed effects, they are identified from variation across criteria within a venture.

— INSERT TABLE 3 ABOUT HERE —

Table 3 reports the results. Both evaluator types respond positively and significantly to the signal: grades rise when the application reads as more positive on a criterion, confirming that the measure captures content the evaluators actually use.The central finding is that the AI responds more strongly than human judges. In column 2, a one-standard-deviation increase in the signal raises the human grade by about 2.3% of the average grade on that criterion and the AI grade by about 3.1%, so the AI's response exceeds the human benchmark by approximately 34%. The gap is somewhat narrower when identification is restricted to variation across ventures within a criterion (Judge $\times$ Criterion fixed effects, column 3, +25%) and wider when restricted to variation across

criteria within a venture (Judge $\times$ Venture fixed effects, column 4, $+86\%$). The result is robust to replacing the embedding-based signal with a simpler measure built from the length of the application answers most similar to the top- and bottom-score descriptions (Table C.2 in Appendix C).

These estimates are reduced-form: equation 2 regresses grades on an empirical measure of each characteristic in the venture description, and the resulting slope is not the weight the model predicts each evaluator places on that characteristic's signal. Appendix B states the additional assumptions under which the reduced-form slope I estimate is informative about that weight and derives the mapping formally. The logic runs as follows. The model's signal is evaluator-specific: each evaluator forms its own private reading of a venture's quality on a given characteristic, so the theoretical object varies at the venture-evaluator-criterion level. The empirical measure, by contrast, is common to both evaluator types and varies only at the venture-criterion level. The two are linked because both are proxies for the same underlying object, the venture's true quality on that characteristic. Under the assumption that the empirical measure's own measurement error is uncorrelated with true quality and with each evaluator's private reading error, the reduced-form slope on this common proxy is proportional to the evaluator signal weight and the ratio of the AI's coefficient to the human's reflects a difference in how each evaluator weights its own private signal. Among the four specifications reported in Table 3, column 3 maps most closely onto this identification argument because the inclusion of judge $\times$ criterion fixed effects absorbs the prior component of grades, leaving the estimated slope to reflect only the signal-driven part of the grade.

The thoeretical framework assumes that an evaluator's grade on a given criterion responds to the signal extracted for that criterion only and not to signals extracted for other criteria (see Appendix A.1). This restriction matches how evaluators actually encounter the venture: applications are organized as separate question blocks, each targeting a different dimension of the venture, and the evaluation form is a fixed list of criteria scored without any cross-criterion aggregation rule (Section 3.1); the same question-by-question structure is preserved in the prompt used to elicit AI

evaluations (Section 3.3). The restriction is nonetheless strong, since it implies that each evaluator processes criteria independently of one another in forming its own assessment, even though the underlying ventures themselves may exhibit genuine cross-criterion correlation in quality. I test the relevance of this restriction by re-estimating Equation 2 controlling for the signals for all fifteen criteria entered jointly. Table C.3 in Appendix C reports the results. The own-criterion signal response coefficients are stable relative to Table 3, while a joint test of the fourteen off-diagonal, other-criteria coefficients fails to reject that they are statistically different from zero. Cross-criterion spillover is therefore not absent, but it does not bias the estimates for own-criterion signal response.

To understand what drives the AI's stronger responsiveness to information, I allow grades to respond separately to the positive component s_{ik}^{+} of the signal, measuring closeness to the high-score description, and the negative component s_{ik}^{-}, measuring closeness to the low-score description.

— INSERT TABLE 4 ABOUT HERE —

Table 4 delivers two findings. First, as a validation of the measure, both human and AI grades respond significantly to both components in the expected direction: grades rise with positive information and fall with negative information. Second, the AI's stronger responsiveness to information is concentrated entirely on the negative side. Human and AI responses to positive content are statistically indistinguishable, whereas the AI responds significantly more strongly than human judges to negative content.

Finally, I allow the response to vary across criteria. Figure 3 reports, for each criterion, the baseline grade of each evaluator type alongside its estimated signal response.

— INSERT FIGURE 3 ABOUT HERE —

Two patterns emerge. The signal response of both evaluators is strongest on the most verifiable criteria (technical development status, regulation, financing, and customer traction), where

the application contains concrete, checkable information, and weakest on the softer, interpretive dimensions such as user pain point, motivation and feasible solution. This pattern is consistent with the prediction of the model: an evaluator's signal weight should be higher whenever the available information pins down the underlying quality with precision, and verifiable criteria are precisely those on which I expect the text to do so. Across criteria, human and AI responses move together, sharing the same sign and a similar magnitude on almost every criterion, so the two evaluator types read the directional content of the application in broadly the same way. The difference is one of degree: the AI responds somewhat more strongly than human judges on nearly every criterion, customer traction being the only exception.

Taken together, these results support the theoretical framework's central hypothesis: the AI places more weight than human judges on the information extracted from the application text, particularly on negative information and on more objective evaluation dimensions, consistent with a higher ratio of signal precision to prior precision for AI than for human evaluators.

4.3. Evaluations and venture performance

The systematic differences documented above raise the question of which evaluator's scores are more informative about what ventures go on to achieve. Because the qualities under evaluation admit no objective benchmark, and human grades are not themselves a ground truth, association with post evaluation performance is the standard against which the two evaluators can ultimately be compared. I relate AI and human scores to three ex post venture outcomes measured from public records in summer 2024: survival, log full-time-equivalent employment, and a binary indicator for any external capital raised, with employment and capital set to zero for ventures that have closed.

A direct comparison is complicated by selection. Human scores determined admission to the program, and program participation plausibly affects subsequent performance, so among admitted ventures a correlation between human scores and outcomes may reflect the treatment effect

of participation rather than the ex ante informativeness of the scores. AI scores, by contrast, are counterfactual and played no role in admission. Following (McKenzie and Sansone, 2019), I estimate the score-outcome relationship separately for admitted and rejected ventures. The cost of the split is that the admitted subsample is small, which limits the precision of the estimates it yields.

For each outcome Y_i, I estimate

$$Y_i = \alpha + \beta_1 AI_i + \beta_2 H_i + \gamma \mathbf{Z}_i + \theta_{b(i)} + \varepsilon_i, \tag{3}$$

where $AI_i = \sum_k X_{ikAI}$ denotes the sum of the AI's scores across the k criteria for venture i, and $H_i = \frac{1}{|J_i|} \sum_{j \in J_i, j \neq AI} \sum_k X_{ikj}$ denotes the average, across the human judges J_i assigned to venture i, of each judge's sum of scores across the k criteria. Both scores are standardized in the pooled sample of ventures, so coefficients measure the outcome response to a one-standard-deviation increase in an evaluator's score. $\mathbf{Z}_i$ collects venture-level application controls (team size, founder education, customer interviews, capital raised prior to application, incorporation status, sector, and business model), and $\theta_{b(i)}$ are batch fixed effects, which absorb differences in time-to-measurement and average venture quality across the 2021–2024 cohorts. I enter the two scores separately and then jointly, comparing their association with each outcome in isolation and in direct competition.

— INSERT TABLE 5 ABOUT HERE —

Table 5 reports the results. Entered on its own, the AI score is positive and significant for survival and employment in both subsamples, and for capital raised among rejected ventures. The human score is significant only among rejected ventures, for survival and employment; it is not significant for capital raised in either subsample. In the joint specifications, the AI coefficient remains positive and significant for survival and employment in both subsamples, and for capital raised among rejected ventures. Among rejected ventures adding the AI score removes the human score significance for survival; for employment the human coefficient remains significant, though smaller than the AI coefficient; for capital raised the human coefficient is not significant in either

subsample or specification. Table C.4 in Appendix C shows that this pattern does not depend on the specific set of controls used, while Figure D.3 in Appendix D reports the nonparametric bootstrap distribution of the point estimates, showing that their sign and significance are robust to sampling variability.

To assess whether the aggregate association of grades with performance holds across most evaluation dimensions or is concentrated in a few, I disaggregate equation (3) to the criterion level,

$$Y_i = \alpha + \beta_1 AI_{ik} + \beta_2 H_{ik} + \gamma \mathbf{Z}_i + \theta_{b(i)} + \varepsilon_i,$$

estimated separately for each criterion k. Tables C.5–C.7 in Appendix C report the results. For employment, both evaluators' associations hold across most criteria among rejected ventures, and to the same extent: the AI coefficient is significant in the joint specification for eleven of the fifteen criteria, and so is the human coefficient. Among admitted ventures, the AI coefficient is significant for five of the fifteen criteria, while the human coefficient is not significant for any. The AI's association with survival is narrower than with employment, holding for seven of the fifteen criteria among rejected ventures. The human coefficient is significant for only three. Among admitted ventures, the AI coefficient is significant for only one criterion, competitive advantage, and the human coefficient for none. For capital raised, few criteria carry a significant association among rejected ventures: the AI coefficient is significant for five, skills and connections, projected sales, VC interest, founders speed, and business model, while the human coefficient is significant for only one, regulation. Among admitted ventures the pattern reverses: the human coefficient is significant for four criteria, while the AI's only for two. Read together, the criterion-level results sharpen the aggregate pattern in Table 5: for survival and employment the AI's scores' association is strong and at least as broad as the human's in every subsample; for capital raised the two evaluators' scores overall association is small and comparable.

The regressions above characterize the association between grades and outcomes at the conditional mean. For innovation and entrepreneurship, however, identifying the ventures that end up at the extremes of the performance distribution, future outliers and future failures alike, is at least as relevant as studying the average performance (Gala and Schwab, 2024). I study this using employment, pooling admitted and rejected ventures and controlling explicitly for admission status, rather than splitting the sample as in the specifications above. Splitting the sample remains the more rigorous approach for isolating the ex ante association of scores with performance from the accelerator's potential treatment effect. I depart from it here for two reasons specific to the tails of the distribution. First, the accelerator's treatment effect is unlikely to affect outliers: participation shapes outcomes most for ventures near the margin of admission, not for the eventual best and worst performers (Åstebro et al., 2025). Second, nonlinear models are more sensitive to the loss of observations that splitting an already modest sample entails. I report the pooled estimates here and the corresponding admitted/rejected splits in Appendix D.

I first estimate the joint score specification of equation (3) by quantile regression at the 25th, 50th, 75th, and 90th percentiles, to examine whether the association between grades and performance holds throughout the outcome distribution.

— INSERT FIGURE 4 ABOUT HERE —

Figure 4 reports the AI and human coefficients at each quantile. Neither coefficient is significant at the 25th percentile; both become significant from the median onward and remain so through the 90th percentile. The AI coefficient exceeds the human coefficient at every quantile, though the gap is not monotonic and the two confidence intervals overlap.

As a second exercise, I split ventures into three segments of the log-employment distribution, below the median, between the median and the 75th percentile, and above the 75th percentile, and estimate a logit model of the probability that a venture's outcome falls in each segment on the AI and human scores, with the same controls as above. Appendix Figure D.4 reports the results: both

coefficients are negative for the below-median segment and positive for the top quartile, and the AI coefficient is larger in magnitude than the human coefficient at both ends, though the confidence intervals overlap throughout. Together, the two exercises suggest that, for employment, the AI's advantage over human judges is not confined to the conditional mean but is strongest at the top of the performance distribution.

I conclude the analysis of the association between evaluators' grades and venture performance with a rank-based exercise. Comparing performance and evalutors' ranking offers a natural way to evaluate an assessor's judgment, and it is particularly well suited to this setting, where selection decisions depend on the ordering of ventures rather than on the level of their scores. Following (Gonzalez-Uribe and Reyes, 2021), I compute, for each batch, the Spearman rank correlation between each evaluator's ranking of ventures by score, defined as the sum of criterion grades, and the ranking of the same ventures separately by total employment and by cumulative capital raised. As with the quantile regression, Figure 5 reports the results pooling admitted and rejected ventures. Appendix Figure D.7 reports the analogous correlations split by admission status.

— INSERT FIGURE 5 ABOUT HERE —

For employment, the AI's rank correlation exceeds the human's in most batches and in the pooled estimate, corroborating the mean-based and distributional evidence above. For capital raised, the pattern reverses: the human's rank correlation exceeds the AI's in most batches, and the pooled human correlation is again the larger of the two. In both panels the two evaluators' confidence intervals overlap substantially, so the batch-level comparison is suggestive rather than conclusive.

Taken together, these results speak to which evaluator's judgment is more strongly associated with venture outcomes. For survival and employment, every exercise points to a stronger association for the AI than for human judges. For capital raised the evidence is more nuanced: the AI is more strongly associated in the mean-based specifications, but the human's advantage in rank

order suggest that human judges might be better at ordering ventures with respect to future capital accumulation.

5. ROBUSTNESS

This section reports two sets of robustness exercises. The first examines whether the paper's central results, the AI's greater responsiveness to criterion-relevant application content and its association with ex post venture performance, are stable across variation in the configuration used to generate the AI evaluations. The second addresses concerns regarding the short-term horizon of performance measures and the possibility of look-ahead bias in the performance regressions.

5.1. AI Configuration

The AI evaluations reported throughout the paper are generated by a single model queried under a single prompt, so that the documented divergence between AI and human evaluators could in principle reflect this particular configuration rather than a general property of AI-based evaluation. Following (Carlson and Burbano, 2026), I test sensitivity along two dimensions of the AI configuration: the prompting strategy and the underlying AI model.

I first generate AI evaluations under six new prompt specifications, holding fixed the underlying model, the criterion-by-criterion scoring rubric, and the structured output schema used to elicit the scores. The new specifications vary only the instructions describing the evaluation task or, in one case, the order in which application material is presented to the model, following the prompt-engineering strategies summarized in Table C.8 and documented in (Carlson and Burbano, 2026). The exact text of each specification is reported in Appendix G.

Figure D.8 in Appendix D reports the estimated AI-human differential in signal responsiveness from the signal-response regression (Equation 2) under each of the seven prompt specifications. The differential is positive and away from zero for every prompt, and the point estimates cluster

around the value obtained under the baseline prompt. Figure 6 reports the corresponding exercise for the association between AI scores and ex post venture performance (Equation 3), separately for admitted and rejected ventures and for each of survival, employment, and capital raised, under both the standalone and the joint AI-and-human specifications.

— INSERT FIGURE 6 ABOUT HERE —

Among rejected ventures, the baseline prompt yields a positive, statistically significant association between AI scores and ex post performance for survival and employment, and a small association for capital raised; each of the six alternative prompt specifications reproduces this same pattern, matching the baseline in sign, magnitude, and significance for all three outcomes. Among admitted ventures, the baseline association is positive for all three outcomes but estimated with far less precision given the smaller subsample; the alternative prompt specifications preserve this sign and this reduced precision.

I next verify the sensitivity of the empirical results to variation in the AI model generating the evaluations. Together with the baseline model Llama by Meta, I consider models developed by DeepSeek, Alibaba (Qwen), Google (Gemma), and Anthropic (Claude), holding the prompt fixed. Table C.9 in Appendix C summarizes the characteristics of the five models. Their release dates range from September 2024 to February 2026, their scale spans two orders of magnitude, and their training cutoffs range from December 2023 to August 2025. They also differ in design and in availability: four are open-weight, while one is accessible only through a closed API.

Figure D.9 in Appendix D reports the estimated AI-human differential in signal responsiveness under each of the five models. DeepSeek reproduces the baseline model's differential, Gemma shows a positive but visibly smaller differential, while for Qwen and Claude the differential cannot be distinguished from a human-level response to the same criterion-relevant content. Figure 7 reports the corresponding performance regressions by model.

— INSERT FIGURE 7 ABOUT HERE —

The association between AI scores and ex post venture performance is positive across all five models, and remains distinguishable from zero for survival and employment among rejected ventures in particular. The association is typically smaller than for the baseline model, with the exception of Claude. Coefficients on capital raised remain small across all models. The most recent model, Claude, does not deliver a visibly stronger association than the baseline model with venture performance, so that greater model capability does not translate into closer alignment between AI scores and realized business outcomes.

Taken together, these results indicate that the prompt used to elicit AI evaluations has little effect on either the AI's differential responsiveness to application content or its association with venture outcomes. The choice of model matters more: the differential in signal responsiveness relative to human judges is not a general property of AI evaluators, present in some models and absent in others, while the association with venture outcomes is more robust across models, though it never exceeds what is documented for the baseline model, even among more recent ones.

5.2. Performance horizon and look ahead bias

The baseline performance regressions measure survival, employment, and capital raised from public records collected in Summer 2024. For more recently evaluated ventures, this measurment may understate realized performance, since employment growth and fundraising typically accrue over several years. In Spring 2026, I repeated the same data collection procedure described in Section 3.2 to construct a second, medium-term measurement of the same three outcomes for the identical set of 579 ventures.

Table C.10 in Appendix C reports the medium-term descriptive statistics. Two additional years of attrition lowered the pooled survival rate from 73% to 62%. Surviving ventures are larger, with mean employment rising from 6.7 to 9 FTEs, and are substantially more likely to have raised any

external capital, from 11% to 29% of the sample; conditional on raising, the mean amount raised is essentially unchanged (€3.5 million versus €3.4 million). I then estimate the performance regression (Equation 3) using these medium-term outcomes.

— INSERT TABLE 6 ABOUT HERE —

Table 6 reports the results. The association of AI scores with survival is larger and more statistically significant at the longer horizon in both subsamples. The association with employment is likewise larger among admitted ventures and, while somewhat smaller among rejected ventures, remains significant at the 1% level in both periods. The association with probability of raising capital increases among rejected ventures, but becomes statistically indistinguishable from zero when controlling for the human score, which is instead strongly associated with capital raised both alone and in the joint specification. The AI scores-performance association is, if anything, no smaller and more precisely estimated at the longer horizon than in the baseline results. The human score seems the stronger correlate of capital raised at this longer horizon, consistent with the pattern already documented at the shorter horizon in Section 4.3.

A further concern for the validity of the AI-performance association is look-ahead bias, the risk that an AI's predictive content reflects overlap between its training data and the outcomes it is asked to predict, rather than genuine inference (Ludwig et al., 2024). This concern is plausible in this setting: the model's training cutoff falls within the window separating venture evaluation from outcome realization in the baseline measurement. The persistence of the AI scores-performance association across outcomes measured in 2024 and in 2026 is itself evidence against this concern: an association generated by such overlap should not survive a repetition of outcomes measurement more than two years after the model's training cutoff.

As additional evidence against this risk, I estimate the performance equation (Equation 3) fully interacted with a dummy for ventures evaluated in the first half of the sample period, reported in Table C.11 in Appendix. If look-ahead bias were at play, the association between AI scores and

performance should be stronger for the oldest ventures, whose time from evaluation to outcome measurement overlaps most with the model's training cutoff. Instead, for employment, the outcome most strongly associated with AI grades, the association is especially strong among the most recently evaluated ventures, those furthest from any possible overlap with the training data cutoff, the opposite of what a look-ahead-bias account would predict. Overall, while this is not a formal test for look-ahead bias, the evidence from the longer-horizon outcomes and from this heterogeneity analysis suggests that it is unlikely that look-ahead bias is a major driver of the association between AI grades and venture performance.

6. DISCUSSION

This paper examines how human experts and AI models evaluate early-stage ventures at a startup accelerator, scoring applications on a set of pre-determined criteria with pre-defined levels. It proposes a conceptual framework relating their divergence to differences in priors and differences in information-extraction precision; it shows that AI evaluations respond more strongly than human evaluations to venture-specific information; and provides evidence that AI scores are at least as strongly associated with ex post venture performance as human scores.

The model I propose is a simple Bayesian-learning model. Bayesian-learning models have a long tradition in strategy and entrepreneurship (Kerr et al., 2014; Gans et al., 2019; Camuffo et al., 2024; Agrawal et al., 2025). Human evaluators are well documented to depart from Bayesian benchmarks in systematic ways (Grether, 1980; Benjamin, 2019), yet simple Bayesian specifications have been shown to meaningfully track observed evaluative judgments (Lakkaraju et al., 2015; Åstebro et al., 2025). Moreover, AI language models are probabilistic models trained to predict the next token given context, a mechanism shown to align closely with Bayesian updating (Xie et al., 2022). This conceptual model therefore can be meanigfully used as a tractable abstraction that mirrors how an evaluator turns available information into an assessment and, by imposing the same structure on both AI and humans, allows to identify two key mechanisms that can drive

divergence between their judgment.

The model implies that if an AI evaluator processes information more precisely, relative to its prior knowledge, than humans do, its evaluations should respond more to venture-specific evidence and rely less on priors. Results for the baseline AI model are consistent with this prediction: AI evaluations respond systematically more to application content than human evaluations, especially for negative information and heterogeneously across criteria. This evidence is subject to two limitations. The comparison between AI and human responsiveness to information rests on the assumption that measurement error in the embedding model's representation of application content is uncorrelated with error in each evaluator's own assessment of the application. Showing that the result holds across several alternative measures of application information, built from different embedding models, different definitions of the signal, and both human- and AI-generated textual descriptions of ventures, would strengthen the credibility of the mechanism while relying less heavily on this strong assumption.

More fundamentally, the content of an application is not randomly assigned, so the relationship between content and grades documented here is a correlation, not a causal effect. Establishing causality would require an experiment that randomly varies the information content of otherwise identical applications and has both AI and human judges evaluate them, an experiment I hope to pursue in future work

AI's stronger responsiveness to information content, relative to humans, is not constant across models: the gap closes for the most recent model, whose grades track human grades far more closely than those of earlier models. Read through the lens of the model, this pattern suggests that more recent models hold stronger priors and rely relatively less on the evidence in front of them than earlier models do. Under this interpretation, greater information-processing capacity does not necessarily lead a model to rely more on available evidence; it may instead lead the model to rely

more on its prior knowledge when forming a judgment. Convergence towards human judgment, however, is not necessarily optimal: an AI evaluator may instead generate more value by preserving its distinctiveness relative to human evaluators, thereby capturing the complementarities available when AI and human judgment are combined (Conti and Messinese, 2025).

More recent models do not significantly outperform earlier ones in their association with venture performance. AI scores are strongly associated with venture performance, often more than human scores for employment, but this association does not appear to improve meaningfully with newer models. This speaks directly to the difficulty of the evaluation task itself: assessing early-stage ventures requires envisioning unrealized future scenarios under Knightian uncertainty, conditions under which AI probabilistic reasoning may struggle (Townsend et al., 2025). A stronger association with venture performance could plausibly be achieved by fine tuning the AI evaluator for this specific task, for example by training it on past evaluations jointly with realized outcomes. It would also be interesting to investigate whether a fully unconstrained evaluator, given no evaluation framework to follow, would outperform a model provided with a consolidated framework, one proven successful for human evaluators (MacMillan et al., 1985), but not designed for an AI model with fundamentally different prior knowledge and information-processing abilities.

This paper's greatest limitation is that it shows AI evaluations are strongly associated with venture performance, but does not identify what information the AI is picking up on that drives this association. Nor does this paper provide an estimate of the treatment effect of using AI, instead of humans, to select ventures. Answering this would require an experiment that randomly assigns ventures to AI or human selection, testing whether full automation changes not only the economic performance of selected ventures, but also their composition along other dimensions, such as innovativeness. These are two open questions I intend to pursue in future research.

These open questions aside, the evidence I provide carries a practical implication for organi-

zations that screen early-stage ventures: AI evaluations are at least as strongly associated with venture performance as human evaluations, and this association is unlikely to deteriorate in the future. A machine can already handle the first screening pass as well as a seasoned judge, at a fraction of the time, money, and effort this requires from humans. This raises a pressing question: where does humans still add the most value?

Three candidate roles stand out. The first is complementarity in judgment itself. When the divergence between AI and human assessments is systematic, the two evaluators are imperfect substitutes, and combining them may recover information that neither recovers alone (Boussioux et al., 2024; Dell'Acqua et al., 2025), though the conditions under which combination outperforms the better of the two remains unclear (Vaccaro et al., 2024).

The second is responsibility for the decision. Screening decisions allocate scarce resources and foreclose opportunities, and legal frameworks governing automated decision-making continue to locate accountability with the human standing behind a decision rather than with the system that produced the recommendation (EU Artificial Intelligence Act, 2024), in part to guard against the adverse outcomes that full delegation would make difficult to arrest.[3] The question is not only legal but one of social acceptance too: evidence on algorithm aversion indicates that the acceptability of a decision depends in part on whether responsibility for it rests with an identifiable human (Bigman and Gray, 2018; Longoni et al., 2019).[4]

The third is mentoring. Advice from experienced practitioners has been shown to meaningfully improve startup outcomes (Chatterji et al., 2019; Sariri, 2025), and substantial evidence indicates that accelerators derive much of their value from mentoring and certification rather than from

[3] In a recent interview with *The Economist*, historian and philosopher Yuval Harari pointed to "existential risks" arising from the legal recognition of organizations run by artificial intelligence systems Harari (2026).

[4] This concern surfaced directly in the field. In conversation, a manager of the accelerator noted that whatever gains automated screening might deliver were offset, in his assessment, by the prospect of complaints and disputes from applicants once decisions were seen to have been made by an AI.

screening alone (Hallen et al., 2020; Gonzalez-Uribe and Leatherbee, 2018). If initial screening can be automated, the expert time it currently absorbs is freed for the activity whose returns are best documented. Understanding how these roles will evolve as AI adoption increases is a natural next step for research on management and the division of labor within organizations.

REFERENCES

Agarwal, N., Moehring, A., and Wolitzky, A. (2025). Designing human-ai collaboration: A sufficient-statistic approach. Technical report, National Bureau of Economic Research.

Agrawal, A., Camuffo, A., Gambardella, A., Gans, J., Scott, E. L., and Stern, S. (2025). *Bayesian Entrepreneurship*. MIT Press, Cambridge, MA.

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press, Boston, MA.

Åstebro, T. and Elhedhli, S. (2006). The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, 52(3):395–409.

Åstebro, T., Funck, A., and Giordano, F. (2025). Cognitive complexity, expertise and prediction accuracy in venture selection. Working paper, HEC Paris.

Åstebro, T., Funck, A., and Giordano, F. (2026). Decomposing venture evaluation: Learning about belief formation. *Strategic Entrepreneurship Journal*, pages 1–36.

Bailey, E., Fehder, D., Floyd, E., Hochberg, Y., and Lee, D. J. (2026). Learning to quit? a multi-year, multi-site field experiment with innovation-driven entrepreneurs. Technical report, National Bureau of Economic Research.

Baley, I. and Veldkamp, L. (2023). Bayesian learning. In *Handbook of economic expectations*, pages 717–748. Elsevier.

Baum, J. A. C. and Silverman, B. S. (2004). Picking winners or building them? alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing*, 19(3):411–436.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In Bernheim, B. D., DellaVigna, S., and Laibson, D., editors, *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 69–186. North-Holland.

Biever, C. (2023). The easy intelligence tests that ai chatbots fail. *Nature*, 619(July):686–689.

Bigman, Y. E. and Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181:21–34.

Bonelli, M. (2026). Data-driven investors. *The Review of Financial Studies*, 39(7):1909–1969.

Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V., and Lakhani, K. R. (2024). The crowdless future? generative ai and creative problem-solving. *Organization Science*, 35(5):1589–1607.

Bryan, K. A. and Gans, J. S. (2026). Training ai for when humans will use it. Technical report, National Bureau of Economic Research.

Camuffo, A., Cordova, A., Gambardella, A., and Spina, C. (2020). A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. *Management Science*, 66(2):564–586.

Camuffo, A., Gambardella, A., Messinese, D., Novelli, E., Paolucci, E., and Spina, C. (2024). A scientific approach to entrepreneurial decision making: Large-scale replication and extension. *Strategic Management Journal*, 45(6):1209–1237.

Cao, S., Jiang, W., Wang, J., and Yang, B. (2024). From man vs. machine to man+ machine: The art and ai of stock analyses. *Journal of Financial Economics*, 160:103910.

Carlson, N. A. and Burbano, V. (2026). The use of llms to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal*, 47(3):699–725.

Chatterji, A., Delecourt, S., Hasan, S., and Koning, R. (2019). When does advice impact startup performance? *Strategic Management Journal*, 40(3):331–356.

Cohen, S., Fehder, D. C., Hochberg, Y. V., and Murray, F. (2019). The design of startup accelerators. *Research Policy*, 48(7):1781–1797.

Cohen, S. L. (2014). Accelerating startups: The seed accelerator phenomenon. *ssrn Journal*, page 1.

Conti, A. and Messinese, D. (2025). The selective tailwind effect of ai on startups: Predictions and anomalies. *Available at SSRN 4958898*.

Csaszar, F. A., Ketkar, H., and Kim, H. (2024). Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. *Strategy Science*, 9(4):322–345.

Csaszar, F. A., Peterson, A., and Wilde, D. (2026). The strategic foresight of llms: Evidence from a fully prospective venture tournament. *arXiv preprint arXiv:2602.01684*.

Damodaran, A. (2009). Valuing young, start-up and growth companies: Estimation issues and valuation challenges. *Available at SSRN 1418687*.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

DeGroot, M. H. (2004). *Optimal statistical decisions*. John Wiley & Sons.

Dell'Acqua, F., Ayoubi, C., Lifshitz, H., Sadun, R., Mollick, E., Mollick, L., Han, Y., Goldman, J., Nair, H., Taub, S., et al. (2025). The cybernetic teammate: A field experiment on generative ai reshaping teamwork and expertise. Technical report, National Bureau of Economic Research.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Doshi, A. R., Bell, J. J., Mirzayev, E., and Vanneste, B. S. (2025). Generative artificial intelligence and evaluating strategic decisions. *Strategic Management Journal*, 46(3):583–610.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 june 2024 laying down harmonised rules on artificial intelligence. 2024/1689.

Fafchamps, M. and Woodruff, C. (2017). Identifying gazelles: Expert panels vs. surveys as a means to identify firms with rapid growth potential. *The World Bank Economic Review*, 31(3):670–686.

Fisher, G. and Neubert, E. (2023). Evaluating ventures fast and slow: Sensemaking, intuition, and deliberation in entrepreneurial resource provision decisions. *Entrepreneurship Theory and Practice*, 47(4):1298–1326.

Franke, N., Gruber, M., Harhoff, D., and Henkel, J. (2008). Venture capitalists' evaluations of start-up teams: Trade-offs, knock-out criteria, and the impact of vc experience. *Entrepreneurship Theory and Practice*, 32(3):459–483.

Gala, K. and Schwab, A. (2024). Stars everywhere: Revealing the prevalence of star performers using empirical data published in entrepreneurship research. *Journal of Business Venturing Insights*, 22:e00492.

Gans, J. S., Stern, S., and Wu, J. (2019). Foundations of entrepreneurial strategy. *Strategic Management Journal*, 40(5):736–756.

Gius, L. (2025). Disagreement predicts startup success: Evidence from venture competitions. *Strategy Science*, 10(2):93–108.

Gompers, P. A., Gornall, W., Kaplan, S. N., and Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, 135(1):169–190.

Gonzalez-Uribe, J. and Leatherbee, M. (2018). The effects of business accelerators on venture performance: Evidence from start-up chile. *The Review of Financial Studies*, 31(4):1566–1603.

Gonzalez-Uribe, J. and Reyes, S. (2021). Identifying and boosting "gazelles": Evidence from business accelerators. *Journal of Financial Economics*, 139:260–287.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics*, 95(3):537–557.

Grumbach, C., N Lane, J., and von Krogh, G. (2026). The mean-variance innovation tradeoff in ai-augmented evaluations. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (26-038):26–038.

Hallen, B. L., Cohen, S. L., and Bingham, C. B. (2020). Do accelerators work? if so, how? *Organization Science*, 31(2):378–414.

Haltiwanger, J., Jarmin, R. S., and Miranda, J. (2013). Who creates jobs? small versus large versus young. *Review of Economics and Statistics*, 95(2):347–361.

Harari, Y. N. (2026). Yuval noah harari on the dangers of an AI future. Interview by Zanny Minton Beddoes, *The Economist*, *The Insider* podcast. Recorded 20 August 2026; published 22 August 2026.

Howell, S. T. (2020). Reducing information frictions in venture capital: The role of new venture competitions. *Journal of Financial Economics*, 136(3):676–694.

Huang, L. and Pearce, J. L. (2015). Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions. *Administrative Science Quarterly*, 60(4):634–670.

Jabarian, B. and Henkel, L. (2025). Voice ai in firms: A natural field experiment on automated job interviews. Working paper SSRN.

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall Series in Experimental Psychology. Prentice-Hall, Englewood Cliffs, NJ.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Kerr, W. R., Nanda, R., and Rhodes-Kropf, M. (2014). Entrepreneurship as experimentation. *Journal of Economic Perspectives*, 28(3):25–48.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Kleinert, S. and Urbig, D. (2026). Startup evaluations with generative artificial intelligence: An exploratory study on early-stage investments and survival predictions by large language models. *Entrepreneurship Theory and Practice*, page 10422587261430320.

Lakkaraju, H., Leskovec, J., Kleinberg, J., and Mullainathan, S. (2015). A Bayesian framework for modeling human evaluations. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 181–189.

Lane, J. N., Boussioux, L., Ayoubi, C., Chen, Y. H., Lin, C., Spens, R., Wagh, P., and Wang, P.-H. (2024). *The narrative AI advantage?: A field experiment on generative AI-augmented evaluations of early-stage innovations*. Harvard Business School Boston, MA.

Lerner, J. and Malmendier, U. (2013). With a little help from my (random) friends: Success and failure in post-business school entrepreneurship. *The Review of Financial Studies*, 26(10):2411–2452.

Li, L., Wu, S., Zhou, J., Xu, K., and Li, X. (2026). When the machine and human disagree: Evaluating the novelty in entrepreneurial funding. In *Academy of Management Proceedings*, volume 2026, page 11831. Academy of Management Valhalla, NY 10595.

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650.

Lopez-Lira, A. and Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2024). Large language models: An applied econometric framework. *Annual Review of Economics*, 18.

Lyonnet, V. and Stern, L. H. (2022). Venture capital (mis) allocation in the age of ai. In *Proceedings of the EUROFIDAI-ESSEC Paris December Finance Meeting*.

Maćkowiak, B., Matějka, F., and Wiederholt, M. (2023). Rational inattention: A review. *Journal of Economic Literature*, 61(1):226–273.

MacMillan, I. C., Siegel, R., and Narasimha, P. N. S. (1985). Criteria used by venture capitalists to evaluate new venture proposals. *Journal of Business Venturing*, 1(1):119–128.

Matějka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298.

McKenzie, D. and Sansone, D. (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in nigeria. *Journal of Development Economics*, 141:102369.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Packard, M. D., Clark, B. B., and Klein, P. G. (2017). Uncertainty types and transitions in the entrepreneurial process. *Organization Science*, 28(5):840–856.

Sariri, A. (2025). The economics of advice: Evidence from start-up mentoring. *Management science*, 71(12):10022–10046.

Scott, E. L., Shu, P., and Lubynsky, R. (2020). Entrepreneurial uncertainty and expert evaluation: An empirical analysis. *Management Science*, 66(11):5170–5193.

Shepherd, D. A. and Majchrzak, A. (2022). Machines augmenting entrepreneurs: Opportunities (and threats) at the nexus of artificial intelligence and entrepreneurship. *Journal of Business Venturing*, 37(4):106227.

Simon, H. A. (1971). Designing organizations for an information-rich world. In Greenberger, M., editor, *Computers, Communication, and the Public Interest*, pages 37–72. Johns Hopkins University Press, Baltimore, MD.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690.

Song, M., Podoynitsyna, K., Van Der Bij, H., and Halman, J. I. M. (2008). Success factors in new ventures: A meta-analysis. *Journal of Product Innovation Management*, 25(1):7–27.

Törnberg, P. (2024). Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

Townsend, D. M., Hunt, R. A., Rady, J., Manocha, P., and Jin, J. H. (2025). Are the futures computable? knightian uncertainty and artificial intelligence. *Academy of Management Review*, 50(2):415–440.

Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303.

Vismara, S., Latifi, G., Meinzinger, L., and Pass, A. (2025). Generative ai-powered venture screening: Can large language models help venture capitalists? *International Review of Financial Analysis*, page 104748.

Wright, M. (2018). *Accelerators: Successful venture creation and growth*. Edward Elgar Publishing.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2022). An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations (ICLR)*.

Yu, S. (2020). How do accelerators impact the performance of high-technology ventures? *Management Science*, 66(2):530–552.

Zacharakis, A. L. and Meyer, G. D. (2000). The potential of actuarial decision models: can they improve the venture capital investment decision? *Journal of Business venturing*, 15(4):323–346.

TABLES

Panel A: General Statistics	
Number of companies	579
HEC graduate	40%
Age	33
Female	46%
Team size	4 (2)
Incorporated	64%
Panel B: Business Model	
Subscription	54%
Marketplace	21%
(e)Commerce	15%
Agency / Training center	4%
Media/Social networks	3%
Mobile App	3%
Panel C: Maturity	
Minimum Viable Product	46%
Prototype	28%
Full working Product	26%
Panel D: Sector	
Software IT	23.7%
Finance Real Estate	12.4%
LifeScience Healthcare	11.4%
Hospitality Education	9.7%
Consulting Professional Services	9.3%
Retail Wholesale Distribution	6.9%
Consumer Goods	6.4%
Media Entertainment Culture	6.4%
Other	13.8%

TABLE 1 Applying ventures
The table presents summary statistics of ventures that pass through the eligibility criteria at the HEC Incubateur and have been evalauted by human judges. Panel A provides general statistics on the applicants, including the total number of applications, the proportion of HEC graduates among founders, the average founder age, the likelihood of having at least one female co-founder, the average team size at the time of application (average number of co-founders in parentheses), and the percentage of ventures already legally registered as incorporated. Panel B displays the distribution of business models among applicants. Panel C categorizes the maturity stage of the applicants' products, from early-stage prototypes to fully developed offerings.

Batch	N Ventures	Survival	FTE		€Mln Funding		
		Mean	Mean	$P75$	$P>0$	Mean$\mid>0$	$P75\mid>0$
2021 Fall	62	56%	10.9	15.5	11%	13	6.2
2021 Summer	49	71%	10.3	12	12%	12.7	2.9
2022 Spring	60	82%	5.9	9	15%	0.7	1
2022 Summer	96	64%	8.2	9	12%	1.2	1.6
2022 Winter	66	74%	10.2	13	23%	2.3	1.7
2023 Spring	72	85%	4.8	6	4%	0.4	0.6
2023 Summer	66	76%	3.9	5	9%	0.7	1
2023 Winter	54	81%	3.5	4	7%	0.6	0.7
2024 Winter	54	72%	3.8	5	7%	1	1
All	579	73%	6.7	8	11%	3.5	2

TABLE 2 Ventures Survival Rates, FTE and Funding

The table presents summary statistics for all ventures evaluated across all batches in the dataset. Column (2) reports the number of ventures evaluated; column (3) shows the survival rate, i.e., the percentage of ventures in each batch still active as of Summer 2024. Columns (4)-(5) shows the average and the 75th percentile of FTEs among surviving ventures. Columns (6)-(8) provide metrics on post application funding in millions of euros: the probability of receiving any funding post-application, the mean funding amount if received, and the funding amount for the top 25% by post application funding in the batch.

Dependent Variable:	X_{ijk}			
Model:	(1)	(2)	(3)	(4)
Variables				
$\Delta_{\text{AI-H}}$(Baseline)	0.3696***	0.4624***		
	(0.0328)	(0.0221)		
Human Signal Response	0.0567***	0.0456***	0.0496***	0.0399***
	(0.0093)	(0.0080)	(0.0078)	(0.0061)
$\Delta_{\text{AI-H}}$(Signal Response)	0.0274***	0.0305***	0.0264***	0.0321***
	(0.0073)	(0.0061)	(0.0065)	(0.0044)
Fixed-effects				
Venture Controls	No	Yes	Yes	Yes
Judge-Criterion (1,095)			Yes	
Judge-Venture (2,212)				Yes
Fit statistics				
Human Baseline	2.179	2.006	2.246	2.726
Observations	33,000	33,000	33,000	33,000
R^2	0.06116	0.10511	0.31998	0.38789

TABLE 3 AI vs. Human Application Information Response
This tables reports OLS estimates from the regression

$$X_{ijk} = \alpha + \alpha_{\text{AI}} D_{j=\text{AI}} + \beta_H s_{ik} + \beta_{AI}(D_{j=\text{AI}} \times s_{ik}) + \\ + \gamma_H \mathbf{Z_i} + \gamma_{AI}(D_{j=\text{AI}} \times \mathbf{Z_i}) + \delta_{jk} + \lambda_{ij} + \varepsilon_{ijk}$$

where $X_{i,j,k}$ is the grade of venture i on criterion k by judge j; s_{ik} measures the direction of information on venture i about criterion k and it is standardized by criterion; $\mathbf{Z_i}$ is a vector of standardized venture-level controls measured at the time of evaluation, including team size and experience, founder education title, gender and years of experience, number of customer interviewed, capital raised, and incorporation status; δ_{jk} are judge–criterion fixed effects, and λ_{ij} are judge–venture fixed effects."Human Signal Response" is β_H and measures how much human grades change for one standard deviation in information content; "$\Delta_{\text{AI-H}}$ (Signal Response)" is the additional AI response β_{AI} and measures whether AI grades change more than human grades for one standard deviation in information content; "$\Delta_{\text{AI-H}}$ (Baseline)" is the AI–human level difference in grades α_{AI}. Column (1) includes no controls; column (2) adds venture controls; column (3) adds judge–criterion fixed effects; column (4) adds judge–venture fixed effects. Standard errors clustered at the judge and venture level in parentheses.Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Dependent Variable:	X_{ijk}			
Model:	(1)	(2)	(3)	(4)
Variables				
$\Delta_{\text{AI-H}}$(Baseline)	0.3695***	0.4606***		
	(0.0329)	(0.0224)		
Human Positive Signal Response	0.0571***	0.0447***	0.0501***	0.0451***
	(0.0097)	(0.0086)	(0.0087)	(0.0071)
$\Delta_{\text{AI-H}}$(Positive Signal Response)	-0.0093	-0.0050	-0.0104	-0.0098**
	(0.0073)	(0.0062)	(0.0072)	(0.0049)
Human Negative Signal Response	-0.0627***	-0.0513***	-0.0564***	-0.0338***
	(0.0097)	(0.0084)	(0.0078)	(0.0067)
$\Delta_{\text{AI-H}}$(Negative Signal Response)	-0.0149**	-0.0175***	-0.0123**	-0.0292***
	(0.0065)	(0.0056)	(0.0052)	(0.0043)
Fixed-effects				
Venture Controls	No	Yes	Yes	Yes
Judge-Criterion (1,095)			Yes	
Judge-Venture (2,212)				Yes
Fit statistics				
Human Baseline	2.179	2.007	2.244	2.720
Observations	33,000	33,000	33,000	33,000
R^2	0.06102	0.10472	0.31978	0.38723

TABLE 4 AI vs. Human Application Information Response, by direction
This tables reports OLS estimates from the regression

$$\begin{aligned} X_{ijk} = \alpha + \alpha_{\text{AI}} D_{j=\text{AI}} + \beta_{H^+} s_{ik}^{+} + \beta_{AI^+}(D_{j=\text{AI}} \times s_{ik}^{+}) + \\ + \beta_{H^-} s_{ik}^{-} + \beta_{AI^-}(D_{j=\text{AI}} \times s_{ik}^{-}) + \\ + \gamma_H \mathbf{Z_i} + \gamma_{AI}(D_{j=\text{AI}} \times \mathbf{Z_i}) + \delta_{jk} + \lambda_{ij} + \varepsilon_{ijk} \end{aligned}$$

where $X_{i,j,k}$ is the grade of venture i on criterion k by judge j; $s_{ik}^{+}(s_{ik}^{-})$ measures the intensity of positive (negative) information on venture i about criterion k and it is standardized by criterion; $\mathbf{Z_i}$is a vector of standardized venture-level controls measured at the time of evaluation, including team size and experience, founder education title, gender and years of experience, number of customer interviewed, capital raised, and incorporation status; δ_{jk} are judge–criterion fixed effects, and λ_{ij} are judge–venture fixed effects."Human Positive (Negative) Signal Response" is $\beta_{H^+}(\beta_{H^-})$ and measures how much human grades change for one standard deviation in positive (negative) information content; "$\Delta_{\text{AI-H}}$ (Positive (Negative) Signal Response)" is the additional AI response $\beta_{AI^+}(\beta_{AI^-})$ and measures whether AI grades change more than human grades for one standard deviation in positive (negative) information content; "$\Delta_{\text{AI-H}}$ (Baseline)" is the AI–human level difference α_{AI}. Column (1) includes no controls; column (2) adds venture controls; column (3) adds judge–criterion fixed effects; column (4) adds judge–venture fixed effects. Standard errors clustered at the judge and venture level in parentheses.Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Admitted			Rejected		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Survival						
AI Score$_i$	0.1262**		0.1318*	0.0779***		0.0682***
	(0.0603)		(0.0680)	(0.0227)		(0.0250)
H Score$_i$		0.0256	-0.0284		0.0552**	0.0289
		(0.0923)	(0.1023)		(0.0244)	(0.0263)
R^2	0.48061	0.43834	0.48159	0.20671	0.19413	0.20904
Panel B: Employment $\log(1+L_i)$						
AI Score$_i$	0.4847***		0.4677***	0.2889***		0.2385***
	(0.1406)		(0.1531)	(0.0440)		(0.0451)
H Score$_i$		0.2777	0.0858		0.2420***	0.1500***
		(0.2515)	(0.2742)		(0.0551)	(0.0567)
R^2	0.58986	0.53968	0.59069	0.29188	0.26740	0.30478
Panel C: Capital $\mathbf{1}(K_i > 0)$						
AI Score$_i$	0.1234		0.0979	0.0234**		0.0178*
	(0.0817)		(0.0835)	(0.0114)		(0.0108)
H Score$_i$		0.1687	0.1286		0.0235	0.0166
		(0.1107)	(0.1063)		(0.0167)	(0.0168)
R^2	0.39802	0.39603	0.40810	0.09745	0.09675	0.09964
Model Specification						
Batch FE	Yes	Yes	Yes	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes	Yes	Yes	Yes
Venture controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	93	93	93	473	473	473

TABLE 5 AI vs. Human Scores and Ventures' Performance
This table reports OLS estimates from the regression

$$Y_i = \alpha + \beta_1 AI_i + \beta_2 H_i + \gamma \mathbf{Z_i} + \theta_{b(i)} + \varepsilon_i.$$

Panel A reports estimates for $Y_i = Survival_i$; Panel B for $Y_i = \log(1+L_i)$, where L_i denotes employment; and Panel C for $Y_i = \mathbf{1}(K_i > 0)$, an indicator for whether the venture raised any capital. Outcomes are constructed from publicly available sources collected in summer 2024. $AI_i = \sum_k X_{ikAI}$ and $H_i = \frac{1}{|J_i|}\sum_{j\in J_i, j\neq \text{AI}} \sum_k X_{ikj}$ denote, respectively, the AI and human evaluation scores for venture i, and are standardized in the pooled sample of ventures. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects. Columns (1)–(3) report estimates for ventures admitted to the accelerator program, while columns (4)–(6) report estimates for ventures that were not admitted. In columns (1) and (4), only AI scores are included; in columns (2) and (5), only human scores are included; in columns (3) and (6), both AI and human scores are included jointly. Heteroskedasticity-robust standard errors are reported in parentheses. Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Admitted			Rejected		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Survival						
AI Score$_i$	0.2385***		0.2695***	0.1013***		0.0864***
	(0.0860)		(0.0907)	(0.0244)		(0.0266)
H Score$_i$		-0.0456	-0.1561		0.0777***	0.0444
		(0.1134)	(0.1255)		(0.0273)	(0.0285)
R^2	0.31517	0.23676	0.33037	0.13333	0.11726	0.13814
Panel B: Employment $\log(1+L_i)$						
AI Score$_i$	0.8856***		0.9692***	0.2387***		0.2133***
	(0.2012)		(0.2118)	(0.0538)		(0.0573)
H Score$_i$		-0.0233	-0.4208		0.1579**	0.0756
		(0.4269)	(0.3915)		(0.0710)	(0.0738)
R^2	0.44643	0.33034	0.45809	0.13408	0.11381	0.13658
Panel C: Capital $\mathbf{1}(K_i > 0)$						
AI Score$_i$	0.0340		0.0089	0.0563**		0.0365
	(0.1009)		(0.1036)	(0.0223)		(0.0232)
H Score$_i$		0.1301	0.1265		0.0728***	0.0587***
		(0.1280)	(0.1310)		(0.0215)	(0.0225)
R^2	0.32620	0.33432	0.33440	0.14732	0.15351	0.15843
Model Specification						
Batch FE	Yes	Yes	Yes	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes	Yes	Yes	Yes
Venture controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	93	93	93	473	473	473

TABLE 6 AI vs. Human Scores and Ventures' Performance, medium term
This table reports OLS estimates from the regression

$$Y_i = \alpha + \beta_1 AI_i + \beta_2 H_i + \gamma \mathbf{Z_i} + \theta_{b(i)} + \varepsilon_i.$$

Panel A reports estimates for $Y_i = Survival_i$; Panel B for $Y_i = \log(1+L_i)$, where L_i denotes employment; and Panel C for $Y_i = \mathbf{1}(K_i > 0)$, an indicator for whether the venture raised any capital. Outcomes are constructed from publicly available sources collected in summer 2026. $AI_i = \sum_k X_{ikAI}$ and $H_i = \frac{1}{|J_i|}\sum_{j\in J_i, j\neq \text{AI}} \sum_k X_{ikj}$ denote, respectively, the AI and human evaluation scores for venture i, and are standardized in the pooled sample of ventures. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects. Columns (1)–(3) report estimates for ventures admitted to the accelerator program, while columns (4)–(6) report estimates for ventures that were not admitted. In columns (1) and (4), only AI scores are included; in columns (2) and (5), only human scores are included; in columns (3) and (6), both AI and human scores are included jointly. Heteroskedasticity-robust standard errors are reported in parentheses. Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

FIGURES

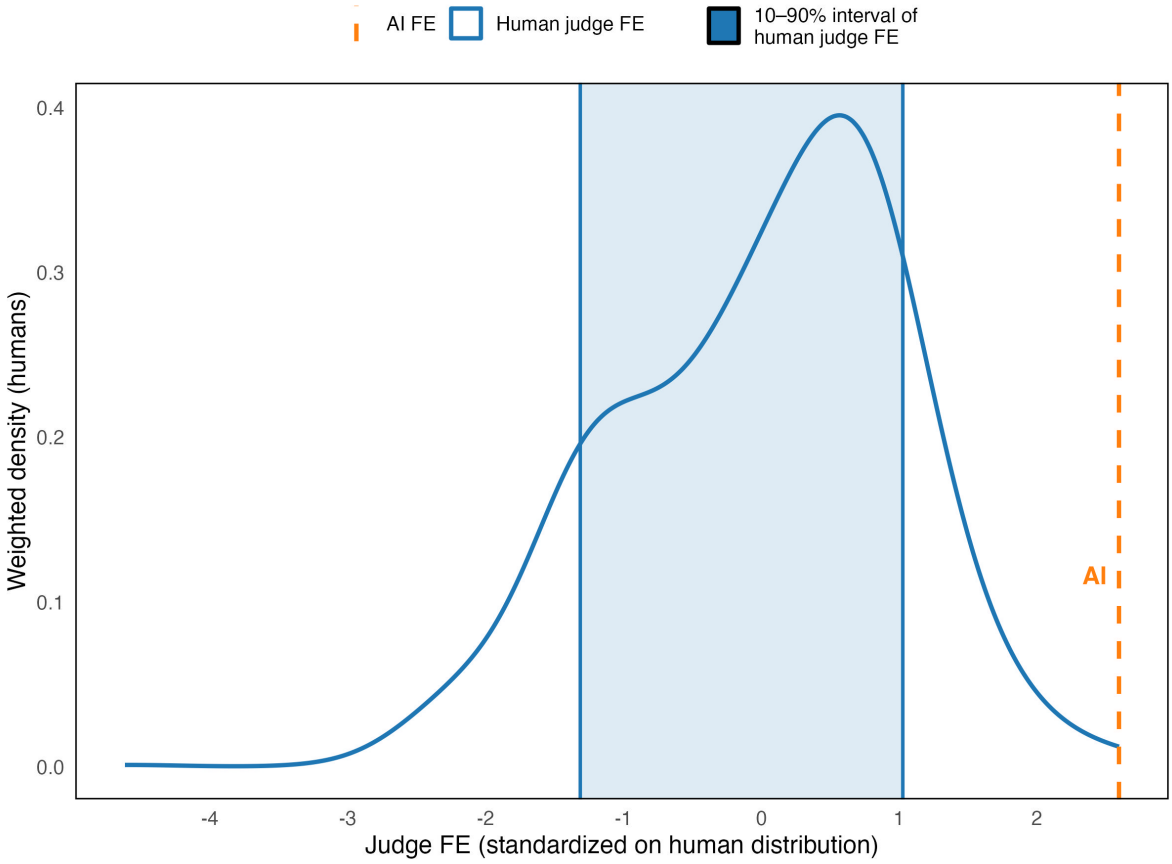

FIGURE 1 AI vs human grades
The figure shows the distribution of standardized judge fixed effects estimated from the two-way fixed effects regression

$$X_{i,j,k} = \alpha_j + \theta_{ik} + \varepsilon_{i,j,k},$$

where $X_{i,j,k}$ is the grade of venture i on criterion k by judge j, α_j is a judge fixed effect, and θ_{ik} is a venture–criterion fixed effect. Fixed effects are standardized so that the human-weighted mean is zero and the human-weighted standard deviation is one, and the density is weighted by the number of evaluations per judge. The vertical orange dashed line reports the AI's fixed effect, indicating whether it is systematically more lenient or stricter than the average human judge. The shaded area corresponds to the 10–90% interval of the human judge fixed effects.

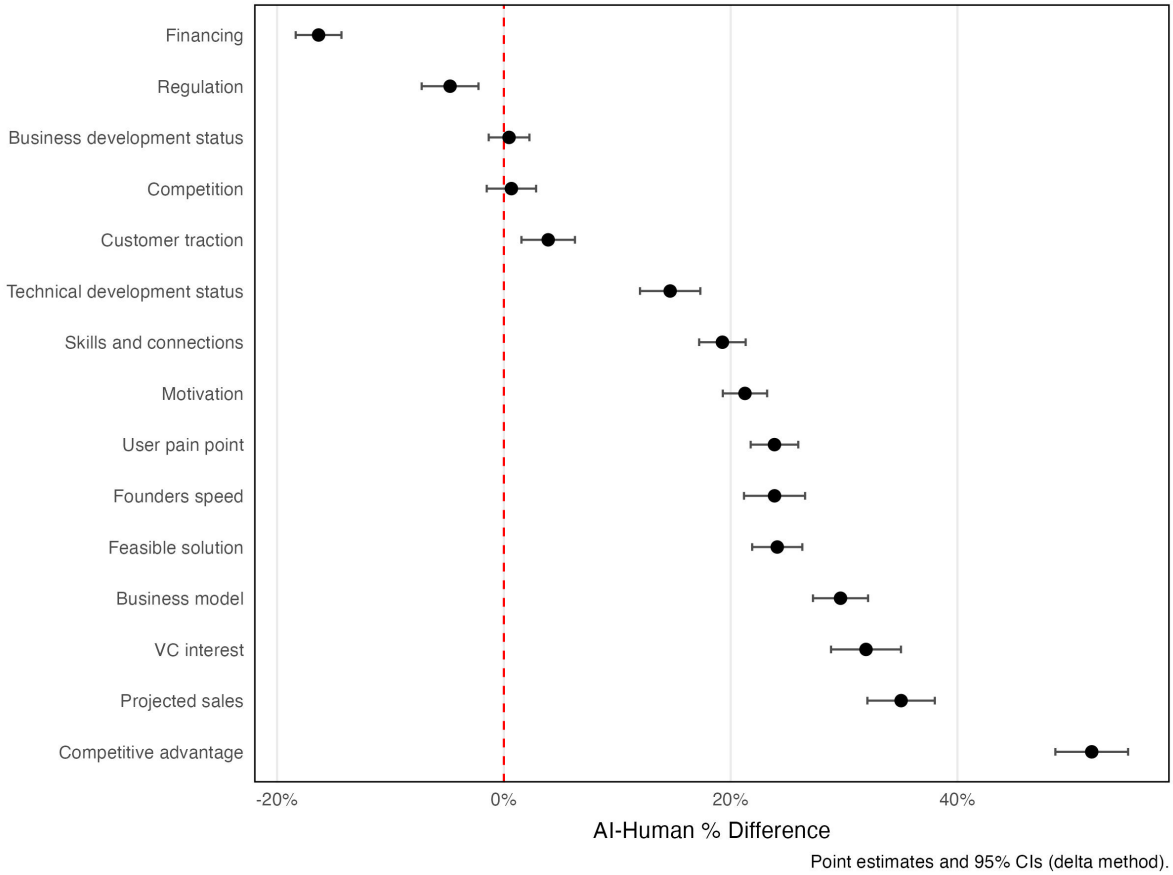

FIGURE 2 AI vs. Human Grades % Difference by criterion
For each criterion $k = 1, \ldots, 15$, the figure reports the point estimate and 95% confidence interval of the relative difference between the AI and human average grade. Estimates are obtained from the regression

$$X_{ijk} = \alpha + \sum_{k=1}^{15} \beta^k D_k + \beta_{\text{AI}} D_{j=\text{AI}} + \sum_{k=1}^{15} \beta_{\text{AI}}^k D_k D_{j=\text{AI}} + \varepsilon_{ijk},$$

where X_{ijk} is the grade of venture i on criterion k by judge j, D_k is a dummy for criterion k, $D_{j=\text{AI}}$ is an indicator for the AI. The reference criterion is $k =$ "Founders speed": for this criterion, the human avarage grade μ_k^H corresponds to the intercept, while the AI average grade μ_k^L corresponds to the intercept plus β_{AI}. For the other criteria, the human average adds the corresponding β^k term, and the AI estimate additionally includes the interaction coefficient β_{AI}^k. The plotted statistic is the percentage difference

$$r_k = \frac{\mu_k^L - \mu_k^H}{\mu_k^H},$$

Standard errors and confidence intervals for r_k are computed by the delta method using the clustered variance–covariance matrix of the regression (clustered at the venture level). This difference range between -15% (Financing) and + 52% (Competitive advantage), with median value +21% (Motivation).

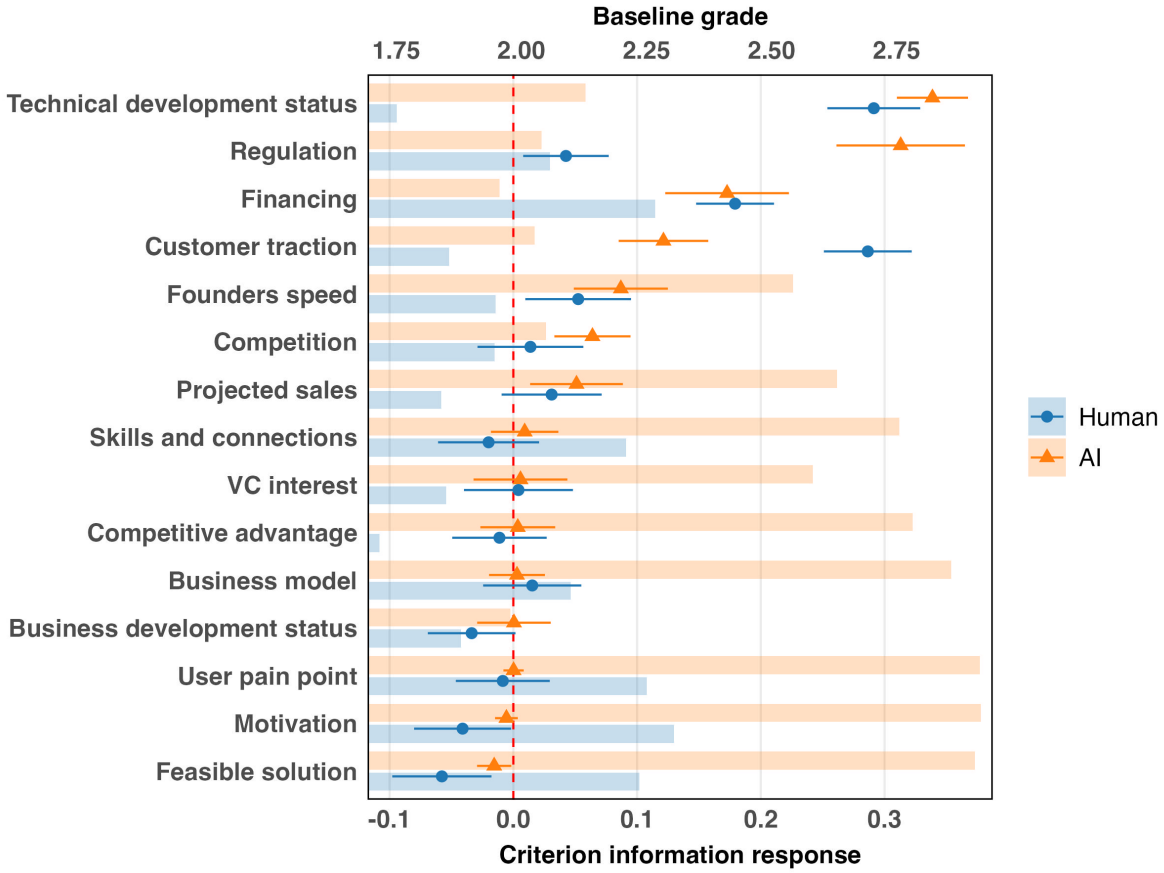

FIGURE 3 AI vs Human Signal Response, by criterion
This figure reports OLS estimates from the regression

$$X_{ijk} = \alpha + \alpha_{\text{AI}} D_{j=\text{AI}} + \beta_H^{14} s_{i,k} + \beta_{AI}^{14} (D_{j=\text{AI}} \times s_{i,k}) + \sum_{k=1}^{15} \beta_H^k (D_k \times s_{i,k}) + \sum_{k=1}^{15} \beta_{AI}^k (D_k \times D_{j=\text{AI}} \times s_{i,k})$$
$$+ \gamma_H \mathbf{Z_i} + \gamma_{AI} (D_{j=\text{AI}} \times \mathbf{Z_i}) + \delta_{jk} + \varepsilon_{ijk}$$

where s_{ik} measures the direction of information on venture i about criterion k and it is standardized by criterion; $\mathbf{Z_i}$ is a vector of standardized venture-level controls measured at the time of evaluation, including team size and experience, founder education title, gender and years of experience, number of customer interviewed, capital raised, and incorporation status; and δ_{jk} are judge×criterion fixed effects. For each criterion k, the figure reports two quantities. The horizontal bars, read against the secondary (top) axis, give the criterion-level baseline grade for human (blue) and AI (orange) evaluators, recovered from the judge×criterion fixed effects δ_{jk}. The point estimates with 95% confidence intervals, read against the primary (bottom) axis, give the criterion-level signal response for one standard deviation in information content for human (blue) and AI (orange) evaluators: the human response on criterion k is the sum of the reference-criterion response and the characteristic-specific term, $\beta_H^{14} + \beta_H^k$, and the AI response is the further sum of the corresponding AI interactions, $\beta_H^{14} + \beta_H^k + \beta_{AI}^{14} + \beta_{AI}^k$. Standard errors are clustered at the venture level.

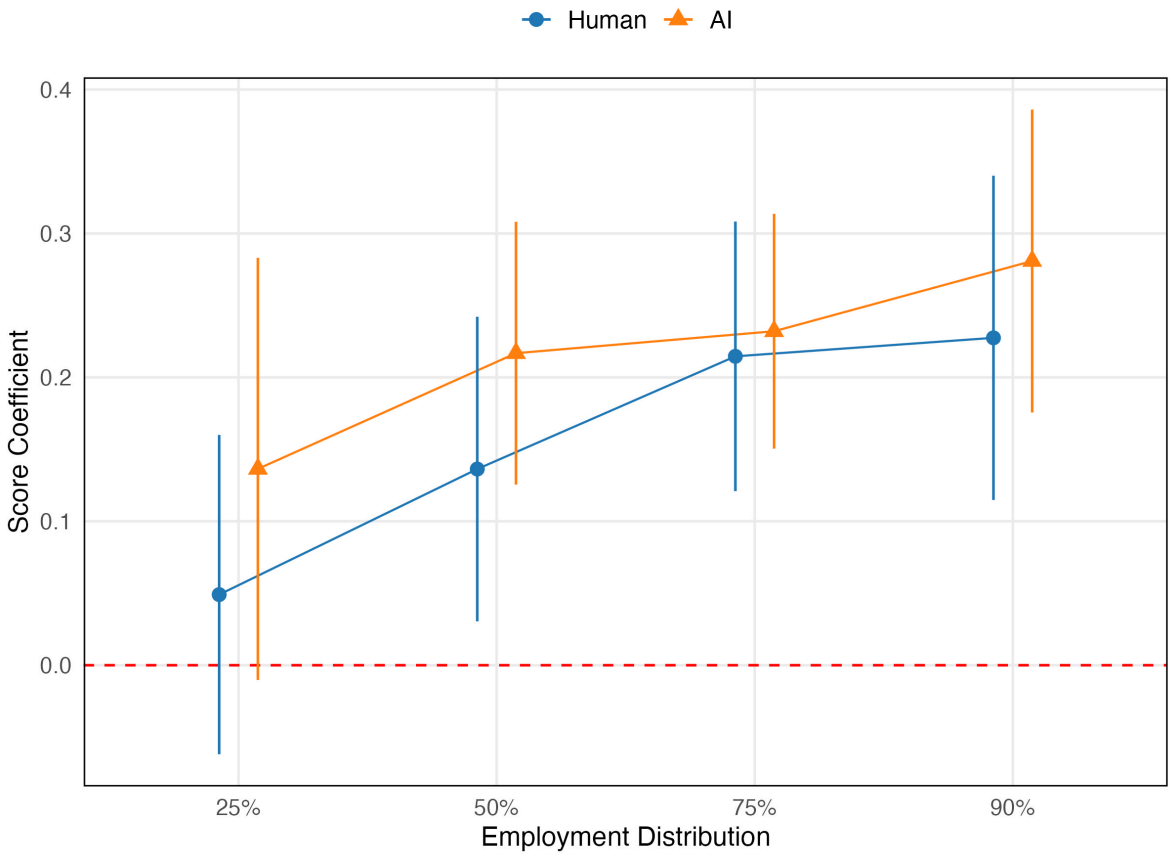

FIGURE 4 AI vs Human Scores and Employment, by quantile
This figure reports estimates from quantile regressions of log employment on AI and human evaluation scores, estimated at the 25th, 50th, 75th, and 90th percentiles, pooling admitted and rejected ventures:

$$\log(1+L_i) = \alpha + \beta_1 AI_i + \beta_2 H_i + \gamma \mathbf{Z_i} + \theta_{b(i)} + \varepsilon_i.$$

$AI_i = \sum_k X_{ik\text{AI}}$ and $H_i = \frac{1}{|J_i|} \sum_{j \in J_i,\, j \neq \text{AI}} \sum_k X_{ikj}$ denote, respectively, the AI and average human evaluation scores for venture i, where J_i is the set of human judges evaluating venture i, and they are standardized in the sample of ventures. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size , number of customer interviews, capital raised, IT sector dummy, and incorporation status; $\theta_{b(i)}$ are batch fixed effects. The specification additionally controls for admission status. The figure plots the point estimates and 95% confidence intervals (heteroskedasticity-robust standard errors) for β_1 (AI, orange) and β_2 (human, blue) at different percentiles of the log-employment distribution.

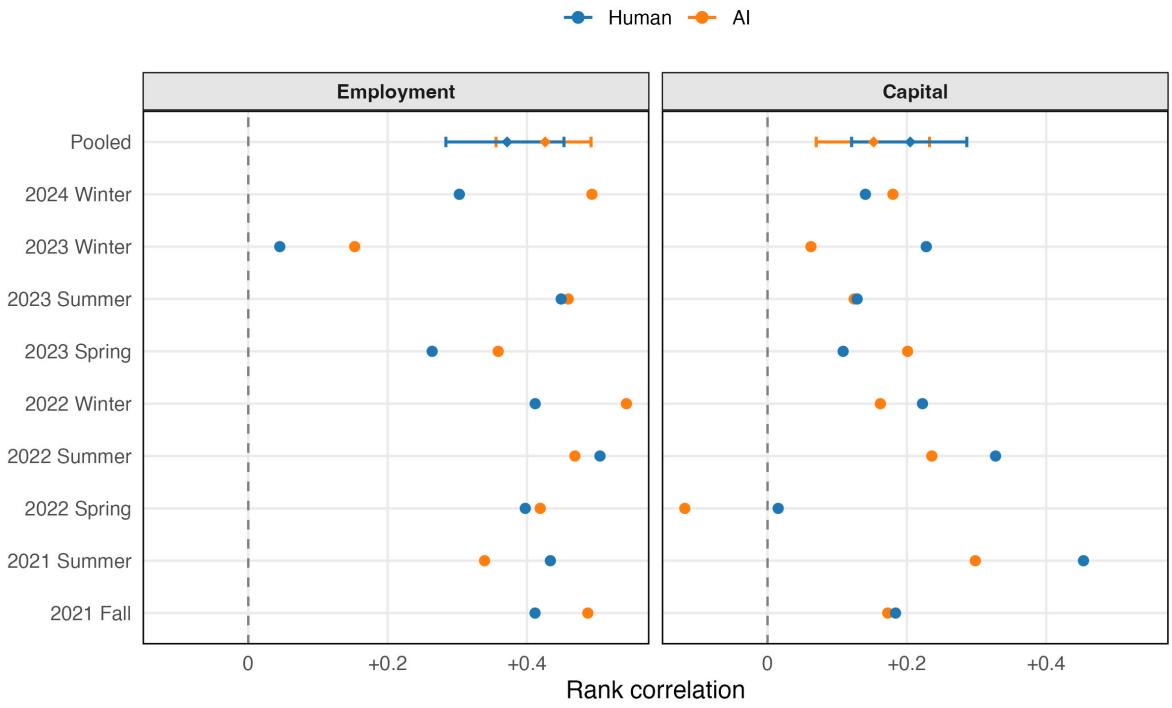

FIGURE 5 Evaluators and Performance Rankings
This figure reports, for each batch, the correlation between each evaluator's ranking of ventures by the score they assigned, defined as the sum of criterion grades, and the realized outcome, separately for employment (left panel) and capital raised (right panel), with admitted and rejected ventures ranked together within each batch. Blue points report the correlation with human ranking; orange points report the correlation with AI ranking. The "Pooled" row aggregates the batch-level correlations via Fisher-z-transformed random-effects (REML) meta-analysis and reports the back-transformed pooled correlation together with its 95% confidence interval. Batch-level correlations are point estimates only, no confidence interval is shown for them individually, since each batch contains too few ventures to estimate one with any precision.

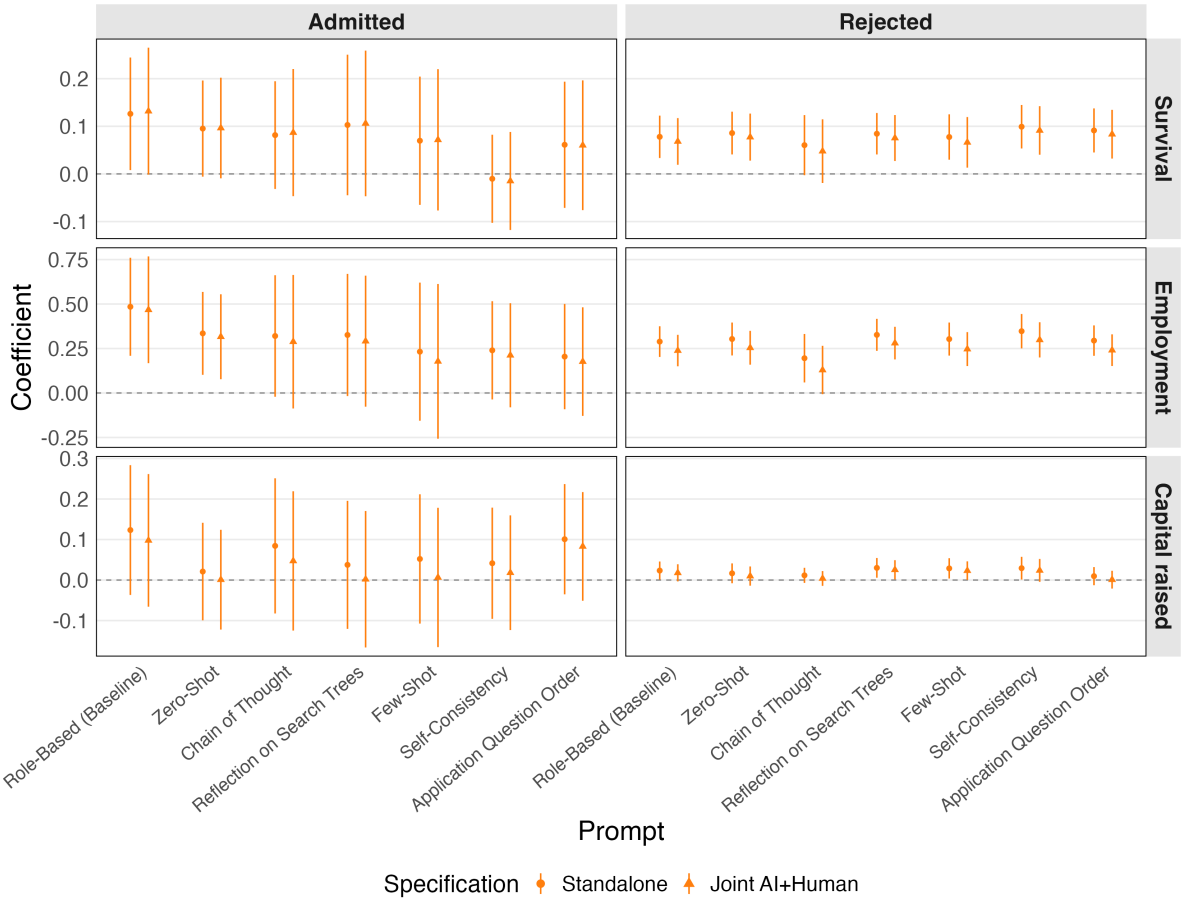

FIGURE 6 AI scores and performance, by prompt
This figure reports coefficients from the regression

$$Y_i = \alpha + \beta_1 AI_i + \beta_2 H_i + \gamma Z_i + \theta_{b(i)} + \varepsilon_i,$$

estimated separately under each of the seven prompt specifications used to generate the AI grades, for admitted and rejected ventures (columns) and for survival, employment, and capital raised (rows). $AI_i = \sum_k X_{ikAI}$ and $H_i = \frac{1}{|J_i|} \sum_{j \in J_i,\, j \neq \text{AI}} \sum_k X_{ikj}$ denote the AI and average human evaluation scores for venture i, standardized in the pooled sample of ventures; coefficients therefore measure the outcome response to a one-standard-deviation increase in an evaluator's score. Z_i is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects. Standalone estimates the AI coefficient with the AI score as the only evaluator score included; Joint AI+Human estimates it jointly with the average human score, both entering the same specification. Vertical lines report 95% confidence intervals from heteroskedasticity-robust standard errors.

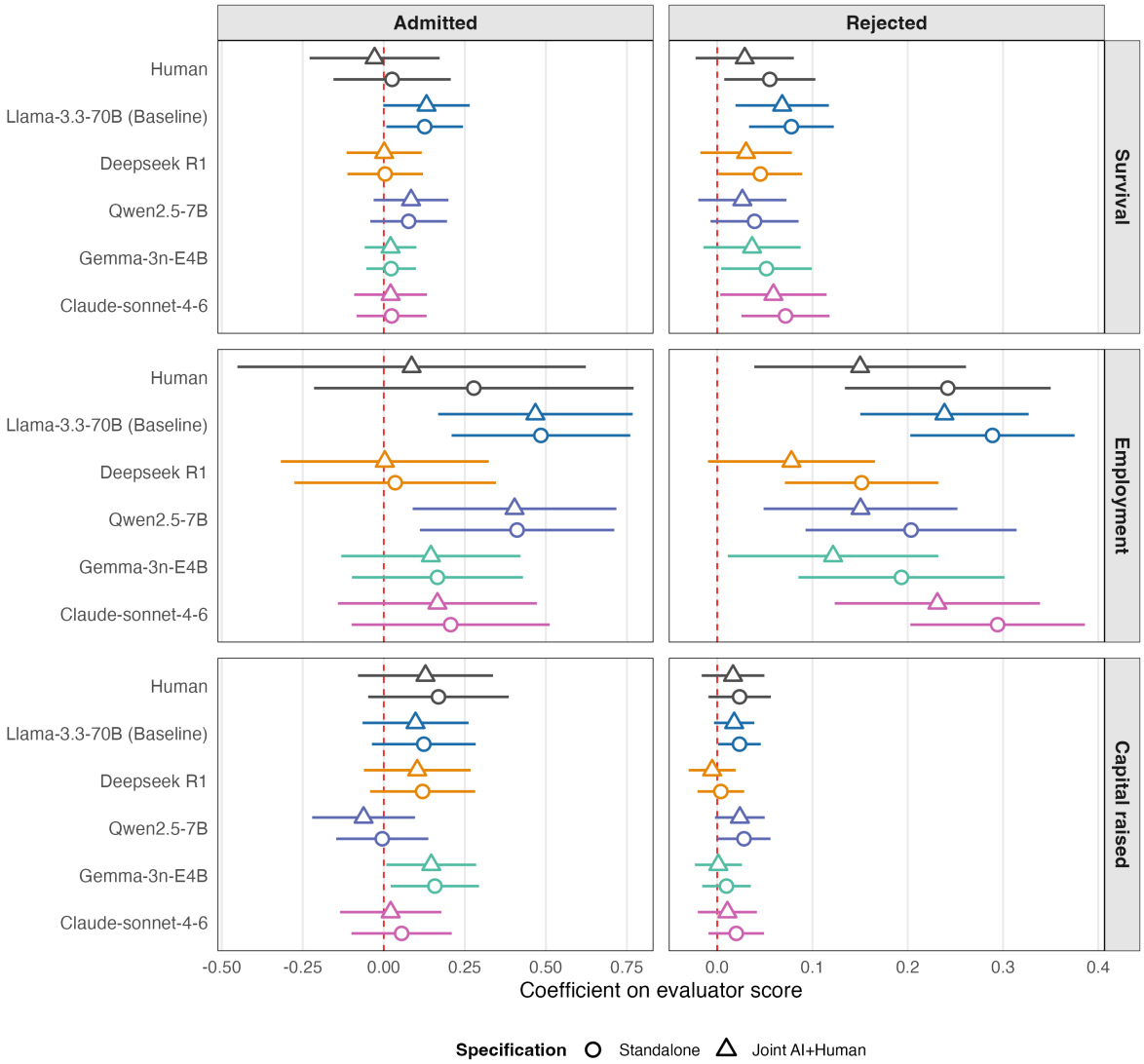

FIGURE 7 AI scores and performance, by AI model
This figure reports coefficients from the OLS regression

$$Y_i = \alpha + \beta_1 AI_i + \beta_2 H_i + \gamma Z_i + \theta_{b(i)} + \varepsilon_i,$$

estimated separately for five different AI models, for admitted and rejected ventures (columns) and for survival, employment, and capital raised (rows). $AI_i = \sum_k X_{ikAI}$ and $H_i = \frac{1}{|J_i|} \sum_{j \in J_i,\, j \neq \text{AI}} \sum_k X_{ikj}$ denote the AI and average human evaluation scores for venture i, standardized in the pooled sample of ventures; coefficients therefore measure the outcome response to a one-standard-deviation increase in an evaluator's score. Z_i is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects.The Human row reports β_2; each AI-model row reports β_1 from the corresponding specification using that model's scores in place of AI_i. Standalone estimates the coefficient with only the evaluator score of interest included; Joint AI+Human estimates it jointly with the other evaluator's score, both entering the same specification. Vertical lines report 95% confidence intervals from heteroskedasticity-robust standard errors.

APPENDIX A: FORMAL MODEL

This appendix presents the formal model in full detail: I first state all assumptions, derivations, and proofs, then present two extensions of the model.

A.1. Assumptions, derivations and proofs

Assumption 1 (Multidimensional latent quality)**.** *Each venture i is characterized by a two-dimensional vector of latent economic qualities,*

$$\mathbf{q}_i \;=\; (q_{i1}, q_{i2})' \;\in\; \mathbb{R}^2,$$

drawn from a jointly Gaussian population distribution

$$\mathbf{q}_i \;\sim\; \mathcal{N}\big(\mathbf{0}, \Sigma_q\big), \qquad \Sigma_q \;=\; \sigma_q^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad \rho \in (-1, 1).$$

The parameter ρ captures the population-level correlation between characteristic-specific qualities. The zero-mean normalization is without loss of generality.

An evaluator, either human or AI, indexed by j, assesses each venture i on each characteristic k. The evaluator chooses a score on each characteristic to come as close as possible to the latent quality on that characteristic, in expectation, given the information the evaluator has processed.

Assumption 2 (Quadratic loss)**.** *Evaluator j chooses a score $m_{ijk} \in \mathbb{R}$ on each characteristic k to minimize expected quadratic loss with target q_{ik}:*

$$m^*_{ijk} \;=\; \arg\min_{m \in \mathbb{R}} \; \mathbb{E}\big[(m - q_{ik})^2 \,\big|\, \mathscr{I}_j(i)\big], \tag{4}$$

where $\mathscr{I}_j(i)$ denotes evaluator j's information set after processing venture i.

Before accessing any specific information, evaluators carry baseline expectations about each characteristic.

Assumption 3 (Gaussian, characteristic-specific, diagonal prior)**.** *Before accessing any information, evaluator j holds independent Gaussian priors over each q_{ik}:*

$$q_{ik} \;\sim\; \mathcal{N}\Big(\mu_{jk},\, \tau_{0jk}^{-1}\Big), \qquad k \in \{1, 2\},$$

where $\mu_{jk} \in \mathbb{R}$ is the prior mean and $\tau_{0jk} > 0$ is the prior precision.

The diagonal-prior (independence across characteristics) restriction implies that each evaluator processes characteristics independently of one another in their own belief system, even though the population of ventures exhibits cross-characteristic correlation through ρ.

When the evaluator accesses information, they extract a noisy reading of the latent quality on each characteristic.

Assumption 4 (Gaussian signals)**.** *Upon accessing information, evaluator j observes a characteristic-specific signal about each latent quality q_{ik}:*

$$s_{ijk} \;=\; q_{ik} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim \mathcal{N}\left(0,\, \tau_{\varepsilon j}^{-1}\right), \qquad k \in \{1,2\}.$$

The signal precision $\tau_{\varepsilon j} > 0$ is evaluator-specific and common across characteristics.

Heterogeneity in $\tau_{\varepsilon j}$ across evaluator types admits a natural microfoundation in the rational-inattention literature (Sims, 2003; Maćkowiak et al., 2023), in which the signal precision is derived as the outcome of an optimal cognitive-resources allocation problem under costly information processing. In that problem, the optimal $\tau_{\varepsilon j}$ is higher when the evaluator's marginal cost of processing information is low relative to the confidence in their prior beliefs: cheaper cognition raises the precision the evaluator finds it worthwhile to acquire, while a more confident prior lowers the marginal benefit of acquiring characteristics-specific information. The formal derivation is provided in Appendix A.2.2.

The posterior characterization developed next also relies on standard regularity conditions governing how signal noise behaves within and across ventures, evaluators, and characteristics. I state these independence conditions explicitly before turning to the model's central result.

Assumption 5 (Independence)**.** *The following independence conditions hold:*

1. $\varepsilon_{ijk} \perp \mathbf{q}_{i'}$ *for all* (i,j,k) *and all* i'*: signal noise is independent of latent quality, including for other ventures.*
2. $\varepsilon_{ijk} \perp \varepsilon_{i'j'k'}$ *for all* $(i,j,k) \neq (i',j',k')$*: noise is independent across evaluator–venture–characteristic triples.*
3. *Each evaluator j observes only their own signal vector $\mathbf{s}_{ij} = (s_{ij1}, s_{ij2})'$, not the signals of other evaluators.*

The first two conditions are the standard signal-extraction assumptions used in normal–normal Bayesian models; they rule out, for instance, common reading biases that would systematically push two evaluators' signals in the same direction for the same venture. The third condition postulates that each judge evaluates applications independently.

With the prior, the signal, and these independence conditions in place, the evaluator's inference problem reduces to normal–normal Bayesian updating applied characteristic by characteristic. I first record, for reference, the standard normal–normal conjugacy result on which the posterior characterization relies.

Lemma 1 (Univariate normal–normal conjugacy)**.** *Let $\theta \sim \mathcal{N}(\mu_0, \tau_0^{-1})$ and let $y \mid \theta \sim \mathcal{N}(\theta, \tau_\varepsilon^{-1})$ with $\tau_0, \tau_\varepsilon > 0$. Then the posterior is*

$$\theta \mid y \;\sim\; \mathcal{N}\left(\mu_1,\, \tau_1^{-1}\right),$$

where

$$\tau_1 \;=\; \tau_0 + \tau_\varepsilon, \qquad \mu_1 \;=\; \frac{\tau_0\,\mu_0 + \tau_\varepsilon\, y}{\tau_0 + \tau_\varepsilon} \;=\; (1-\omega)\,\mu_0 + \omega\, y, \qquad \omega \;=\; \frac{\tau_\varepsilon}{\tau_0 + \tau_\varepsilon}.$$

Proof. By Bayes' rule, $p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta)$. Substituting the two Gaussian densities and collecting terms in θ in the exponent:

$$\begin{aligned} p(\theta \mid y) &\propto \exp\!\left[-\tfrac{1}{2}\tau_\varepsilon (y-\theta)^2\right] \cdot \exp\!\left[-\tfrac{1}{2}\tau_0(\theta-\mu_0)^2\right] \\ &= \exp\!\left[-\tfrac{1}{2}\Big((\tau_0+\tau_\varepsilon)\theta^2 - 2(\tau_0\mu_0+\tau_\varepsilon y)\theta + (\tau_\varepsilon y^2 + \tau_0\mu_0^2)\Big)\right] \\ &= \exp\!\left[-\tfrac{1}{2}(\tau_0+\tau_\varepsilon)\left(\theta^2 - 2\frac{\tau_0\mu_0+\tau_\varepsilon y}{\tau_0+\tau_\varepsilon}\,\theta\right)\right] \cdot \exp\!\left[-\frac{1}{2}\left(\tau_\varepsilon y^2 + \tau_0\mu_0^2\right)\right] \\ &= \exp\!\left[-\tfrac{1}{2}(\tau_0+\tau_\varepsilon)\,(\theta-\mu_1)^2\right] \cdot \exp\!\left[-\frac{1}{2}\left(\tau_\varepsilon y^2 + \tau_0\mu_0^2 - (\tau_0+\tau_\varepsilon)\mu_1^2\right)\right] \\ &\propto \exp\!\left[-\tfrac{1}{2}(\tau_0+\tau_\varepsilon)\,(\theta-\mu_1)^2\right] \end{aligned}$$

which is the kernel of a normal distribution with precision $\tau_1 = \tau_0 + \tau_\varepsilon$ and mean μ_1. □

The posterior distribution is Gaussian, and its mean and precision admit closed-form expressions.

Lemma 2 (Posterior distribution)**.** *Under Assumptions 1–5, the posterior distribution of q_{ik} given the signal vector $\mathbf{s}_{ij}$ depends only on the own-characteristic signal s_{ijk} and is Gaussian: $q_{ik} \,\big|\, \mathbf{s}_{ij} \sim \mathcal{N}\left(\mu_{ijk}^{\text{post}},\, \tau_{1jk}^{-1}\right)$, with posterior mean*

$$\mu_{ijk}^{\text{post}} \;=\; (1-\omega_{jk})\,\mu_{jk} + \omega_{jk}\, s_{ijk}, \tag{5}$$

posterior precision $\tau_{1jk} = \tau_{0jk} + \tau_{\varepsilon j}$, and signal weight

$$\omega_{jk} \;\equiv\; \frac{\tau_{\varepsilon j}}{\tau_{0jk} + \tau_{\varepsilon j}} \;\in\; (0,1).$$

Proof. By Assumption 3, the prior on $\mathbf{q}_i$ in the evaluator's belief system factorizes: $p_j(\mathbf{q}_i) = p_j(q_{i1})\, p_j(q_{i2})$. By Assumption 5, signals are conditionally independent across characteristics given $\mathbf{q}_i$: $p(\mathbf{s}_{ij} \mid \mathbf{q}_i) = p(s_{ij1} \mid q_{i1})\, p(s_{ij2} \mid q_{i2})$. Therefore the joint posterior factorizes:

$$p_j(\mathbf{q}_i \mid \mathbf{s}_{ij}) \;\propto\; \prod_{k=1}^{2} p_j(q_{ik})\, p(s_{ijk} \mid q_{ik}),$$

and each marginal posterior is the univariate normal–normal conjugate update. With prior $q_{ik} \sim \mathcal{N}(\mu_{jk}, \tau_{0jk}^{-1})$ and signal $s_{ijk} \mid q_{ik} \sim \mathcal{N}(q_{ik}, \tau_{\varepsilon j}^{-1})$, Lemma 1 gives (5) and the posterior precision $\tau_{1jk} = \tau_{0jk} + \tau_{\varepsilon j}$. The cross-characteristic signal $s_{ijk'}$ for $k' \neq k$ does not enter the marginal posterior on q_{ik} because, under the diagonal prior, q_{ik} and $q_{ik'}$ are independent in the evaluator's belief system and signals are conditionally independent across characteristics. □

The posterior mean is a precision-weighted average of the evaluator's prior mean μ_{jk} and the characteristic-specific signal s_{ijk}, with the weight ω_{jk} on the signal determined by the relative precision of the two inputs. Combining the posterior with the quadratic-loss objective yields the central result of the model.

Proposition 1 (Optimal grade)**.** *Under Assumptions 1–5, the optimal score on characteristic k is*

$$\boxed{m^*_{ijk} \;=\; \underbrace{(1-\omega_{jk})\,\mu_{jk}}_{\text{prior component}} \;+\; \underbrace{\omega_{jk}\,s_{ijk}}_{\text{signal response}}} \tag{6}$$

with ω_{jk} given by Lemma 2.

Proof. The evaluator solves problem (4):

$$m^*_{ijk} \;=\; \arg\min_{m\in\mathbb{R}} \; \mathbb{E}\big[(m-q_{ik})^2 \,\big|\, \mathbf{s}_{ij}\big]\,.$$

Expanding the objective,

$$\mathbb{E}\big[(m-q_{ik})^2 \;\big|\; \mathbf{s}_{ij}\big] = m^2 - 2m\,\mathbb{E}[q_{ik} \mid \mathbf{s}_{ij}] + \mathbb{E}[q_{ik}^2 \mid \mathbf{s}_{ij}],$$

the last term is independent of m. The first-order condition yields $m^*_{ijk} = \mathbb{E}[q_{ik} \mid \mathbf{s}_{ij}] = \mu^{\text{post}}_{ijk}$, and the second-order condition is satisfied since the objective is strictly convex in m. Substituting the expression for the posterior mean from Lemma 2 gives (6). □

A.2. Extensions

This part of the appendix collects two extensions of the model. Section A.2.1 shows why the extension to $K > 2$ characteristics follows mechanically from the model's assumptions. Section A.2.2 provides a rational-inattention microfoundation for the evaluator-specific signal precision $\tau_{\varepsilon j}$.

A.2.1. Extension to $K \geq 2$ Characteristics

The model extends to arbitrary $K \geq 2$ mechanically. Latent quality becomes $\mathbf{q}_i \in \mathbb{R}^K$, drawn from a jointly Gaussian population distribution with an unrestricted positive-definite covariance matrix Σ_q:

$$\mathbf{q}_i \;\sim\; \mathcal{N}\big(\mathbf{0}, \Sigma_q\big)\,, \qquad \Sigma_q \succ 0.$$

The cross-characteristic correlation structure in the population of ventures is left unrestricted. In the evaluator's belief system, the diagonal-prior restriction is maintained: $\mathbf{T}_{0j} = \operatorname{diag}(\tau_{0j1}, \ldots, \tau_{0jK})$. This restriction, and not the structure of Σ_q, is what drives the characteristic-by-characteristic decomposition of the optimal grade: under characteristic-specific Gaussian signals $s_{ijk} = q_{ik} + \varepsilon_{ijk}$ with $\varepsilon_{ijk} \sim \mathcal{N}(0, \tau_{\varepsilon j}^{-1})$, the posterior on each q_{ik} depends only on the own-characteristic signal s_{ijk}, and the optimal grade (6) obtains characteristic by characteristic exactly as in the two-characteristic case.

A.2.2. Signal precision microfoundation

The baseline model treats signal precision $\tau_{\varepsilon j}$ as an exogenous evaluator-specific parameter. This leaves open the source of cross-evaluator differences in precision. This subsection provides a microfoundation in which evaluator j optimally chooses how much cognitive resources to allocate

to the information extraction, trading off the benefit of more precise information against its processing cost (Sims, 2003; Maćkowiak et al., 2023). The key insight is that $\tau_{\varepsilon j}$ can be interpreted as the outcome of an optimal cognitive-resources allocation problem. In the resulting optimum, $\tau_{\varepsilon j}$ is decreasing in the marginal cost of information processing κ_j and decreasing in the precision of the evaluator's priors τ_{0jk}: a more confident prior reduces the marginal benefit of acquiring characteristic-specific information, and costlier cognition raises its marginal cost. Differences in κ_j and τ_{0j} between human experts and the AI model therefore translate into differences in signal precision in the baseline model. Below, I derive this result explicitly, showing in closed form how optimal signal precision depends on the evaluator's cognitive cost and prior precision.

Information Cost

Assumption 6 (Costly cognition)**.** *Before assigning grades, evaluator j chooses the information structure of signals $\mathbf{s}_{ij}$ about $\mathbf{q}_i$. Acquiring more information is costly. Following the rational-inattention literature, I adopt the entropy-based cost function*

$$C_j \;=\; \kappa_j I(\mathbf{q}_i\,;\mathbf{s}_{ij}), \tag{7}$$

where $\kappa_j > 0$ is evaluator j's marginal cost of processing information, and $I(\mathbf{q}_i\,;\mathbf{s}_{ij}) \equiv H(\mathbf{q}_i) - \mathbb{E}\big[H(\mathbf{q}_i \mid \mathbf{s}_{ij})\big]$ is the Shannon mutual information between $\mathbf{q}_i$ and $\mathbf{s}_{ij}$, measuring the expected reduction in uncertainty about venture quality upon observing the signal.

Mutual Information for the Gaussian Signal

Lemma 3 (Mutual information under Gaussian conjugacy)**.** *Under Assumptions 3–5, the mutual information between q_{ik} and the signal s_{ijk} is*

$$I(q_{ik}\,;s_{ijk}) \;=\; \tfrac{1}{2}\log\!\left(1+\frac{\tau_{\varepsilon j}}{\tau_{0jk}}\right),$$

and the total mutual information across characteristics is the sum of the characteristic-specific terms.

Proof. For $X \sim \mathcal{N}(\mu,\sigma^2)$, the differential entropy is $H(X) = \frac{1}{2}\log(2\pi e\sigma^2)$. Under the Gaussian prior and signal, Lemma 1 gives the posterior precision $\tau_{1jk} = \tau_{0jk} + \tau_{\varepsilon j}$, hence

$$I(q_{ik}\,;s_{ijk}) \;=\; \tfrac{1}{2}\log\!\left(2\pi e\,\tau_{0jk}^{-1}\right) - \tfrac{1}{2}\log\!\left(2\pi e\,\tau_{1jk}^{-1}\right) \;=\; \tfrac{1}{2}\log\!\left(1+\tfrac{\tau_{\varepsilon j}}{\tau_{0jk}}\right).$$

Additivity across characteristics follows from the conditional independence of signals (Assumption 5(ii)) together with the diagonal prior (Assumption 3). □

Ex Ante Expected Loss under Optimal Grading

Lemma 4 (Ex ante expected loss)**.** *Given signal precision $\tau_{\varepsilon j}$, the ex ante expected quadratic loss on characteristic k under the optimal grade (Proposition 1) is*

$$\mathbb{E}\Big[(m^*_{ijk} - q_{ik})^2\Big] \;=\; \frac{1}{\tau_{0jk}+\tau_{\varepsilon j}}.$$

Proof. Under quadratic loss, $m^*_{ijk} = \mathbb{E}[q_{ik} \mid s_{ijk}]$. By the law of iterated expectations,

$$\mathbb{E}\Big[(m^*_{ijk} - q_{ik})^2\Big] \;=\; \mathbb{E}\big[(q_{ik} - \mathbb{E}[q_{ik} \mid s_{ijk}])^2\big] \;=\; \mathbb{E}\big[\mathrm{Var}(q_{ik} \mid s_{ijk})\big] \;=\; \tau_{1jk}^{-1},$$

where the last equality uses Lemma 2 and the fact that, in the Gaussian conjugate model, the posterior variance is constant in the realized signal. □

The Cognitive-Resources Allocation Problem The evaluator chooses $\tau_{\varepsilon j} \geq 0$ to minimize the sum of ex-ante expected scoring losses across characteristics plus the cognition cost (Assumption 6):

$$\min_{\tau_{\varepsilon j} \geq 0} \; \underbrace{\sum_{k=1}^{2} \frac{1}{\tau_{0jk} + \tau_{\varepsilon j}}}_{\text{expected grading error}} \; + \; \underbrace{\frac{\kappa_j}{2} \sum_{k=1}^{2} \log\left(\frac{\tau_{0jk} + \tau_{\varepsilon j}}{\tau_{0jk}}\right)}_{\text{cognition cost}}. \tag{8}$$

Proposition 2 (Optimal signal precision)**.** *The first-order condition for problem (8) is*

$$\sum_{k=1}^{2} \frac{1}{(\tau_{0jk} + \tau_{\varepsilon j})^2} \;=\; \frac{\kappa_j}{2} \sum_{k=1}^{2} \frac{1}{\tau_{0jk} + \tau_{\varepsilon j}}. \tag{9}$$

The problem admits a unique interior or boundary solution $\tau^*_{\varepsilon j} \geq 0$. In the symmetric case $\tau_{0j1} = \tau_{0j2} \equiv \tau_{0j}$, the solution takes the closed form

$$\tau^*_{\varepsilon j} \;=\; \max\left\{0,\; \frac{2}{\kappa_j} - \tau_{0j}\right\}. \tag{10}$$

Proof. Differentiating (8) with respect to $\tau_{\varepsilon j}$ yields (9). Let $f(\tau_{\varepsilon j}) \equiv \sum_{k=1}^{2} (\tau_{0jk} + \tau_{\varepsilon j})^{-1}$ and $g(\tau_{\varepsilon j}) \equiv \frac{\kappa_j}{2} \sum_{k=1}^{2} \log[(\tau_{0jk} + \tau_{\varepsilon j})/\tau_{0jk}]$. The second derivative of the objective is

$$\sum_{k=1}^{2} \frac{2}{(\tau_{0jk} + \tau_{\varepsilon j})^3} \;-\; \frac{\kappa_j}{2} \sum_{k=1}^{2} \frac{1}{(\tau_{0jk} + \tau_{\varepsilon j})^2}.$$

At any interior critical point, (9) implies $\frac{\kappa_j}{2} = \sum_k (\tau_{0jk} + \tau_{\varepsilon j})^{-2} / \sum_k (\tau_{0jk} + \tau_{\varepsilon j})^{-1}$. Substituting and applying Cauchy–Schwarz with $a_k = (\tau_{0jk} + \tau)^{-3/2}$ and $b_k = (\tau_{0jk} + \tau)^{-1/2}$ shows that the second derivative is strictly positive at any interior solution. Hence the solution is a strict local minimum, and continuity on $[0, \infty)$ together with the boundary behavior as $\tau \to \infty$ implies uniqueness. In the symmetric case, the FOC reduces to $\tau_{0j} + \tau_{\varepsilon j} = 2/\kappa_j$, with the non-negativity constraint binding whenever $\tau_{0j} \geq 2/\kappa_j$. □

Remark 1 (Asymmetric case)**.** *When prior precisions differ across characteristics, the optimum equates a weighted average of posterior* variances*, $\tau_{1jk}^{-1} = (\tau_{0jk} + \tau_{\varepsilon j})^{-1}$, weighted by those same variances, to the marginal cost of cognition $\kappa_j/2$. By the implicit function theorem applied to (9), the optimum is decreasing in each τ_{0jk} and in κ_j: an evaluator with a more confident prior on any single characteristic, or with more expensive cognition, acquires less information overall.*

Two comparative statics summarize the economics of this allocation problem.

1. *Monotonicity in cognition cost.*

$$\frac{\partial \tau^*_{\varepsilon j}}{\partial \kappa_j} < 0,$$

$\tau^*_{\varepsilon j}$ is decreasing in the marginal cost of cognition κ_j: an evaluator who finds it more costly to extract an additional unit of information optimally chooses to extract less, holding beliefs closer to the prior.

2. *Monotonicity in priors.*

$$\frac{\partial \tau^*_{\varepsilon j}}{\partial \tau_{0jk}} < 0 \qquad \text{for each } k,$$

$\tau^*_{\varepsilon j}$ is decreasing in prior precision τ_{0jk}: an evaluator who already holds a confident view of a characteristic has less to gain from costly information acquisition, since the expected reduction in grading error from a given increase in signal precision is smaller when the prior is already tight.

APPENDIX B: IDENTIFICATION

This appendix states the additional assumptions under which the coefficient estimates on s_{ik} reported in Table 3 of section 4 are informative about the model's signal weight ω_{jk} characterized in Proposition 1.

Assumption 7 (Measurement model for the empirical signal)**.** *For every venture i and criterion k,*

$$s_{ik} = q_{ik} + \eta_{ik}, \qquad \mathbb{E}[\eta_{ik}] = 0, \qquad 0 < \mathrm{Var}(\eta_{ik}) < \infty,$$

where η_{ik} is the extraction error of the embedding-based measurement procedure applied to venture i's information on criterion k.

Assumption 8 (Signal-noise orthogonality)**.** *For every venture i, criterion k, and evaluator $j \in \{H, AI\}$:*

(i) $\mathrm{Cov}(q_{ik}, \eta_{ik}) = 0$*;*

(ii) $\mathrm{Cov}(\varepsilon_{ijk}, \eta_{ik}) = 0$.

Condition (i) requires that the extraction error of the embedding procedure is not systematically larger, or systematically signed, for stronger or weaker ventures. Condition (ii) requires that an individual evaluator's own reading error on a given application is uncorrelated with the embedding procedure's reading error on the same text.

Proposition 3 (Consistency of the reduced-form slope)**.** *Under Assumptions 1–6 and Assumptions 7 and 8 above, the population least-squares slope of X_{ijk} on s_{ik} within the (j,k) cell,*

$$\beta_{jk} \equiv \frac{\mathrm{Cov}(X_{ijk}, s_{ik} \mid j,k)}{\mathrm{Var}(s_{ik} \mid j,k)},$$

satisfies

$$\beta_{jk} = \omega_{jk}\,\lambda_k, \qquad \lambda_k \equiv \frac{\mathrm{Var}(q_{ik})}{\mathrm{Var}(q_{ik}) + \mathrm{Var}(\eta_{ik})} \in (0,1),$$

where λ_k does not depend on j.

Proof. Throughout, $X_{ijk} = m^*_{ijk}$, and all covariances and variances are conditional on (j,k).

Step 1 (Covariance between the grade and the empirical signal). By Proposition 1 and Assumption 4 ($s_{ijk} = q_{ik} + \varepsilon_{ijk}$),

$$X_{ijk} = (1 - \omega_{jk})\mu_{jk} + \omega_{jk}\,q_{ik} + \omega_{jk}\,\varepsilon_{ijk}.$$

By Assumption 7, $s_{ik} = q_{ik} + \eta_{ik}$. Since $(1-\omega_{jk})\mu_{jk}$ is constant within the (j,k) cell (Assumption 2.2), it has zero covariance with s_{ik}, and bilinearity of covariance gives

$$\mathrm{Cov}(X_{ijk}, s_{ik}) = \omega_{jk}\,\mathrm{Cov}\left(q_{ik},\, q_{ik} + \eta_{ik}\right) + \omega_{jk}\,\mathrm{Cov}\left(\varepsilon_{ijk},\, q_{ik} + \eta_{ik}\right).$$

Expanding both covariances,

$$\mathrm{Cov}(X_{ijk}, s_{ik}) = \omega_{jk}\Big[\mathrm{Var}(q_{ik}) + \mathrm{Cov}(q_{ik}, \eta_{ik})\Big] + \omega_{jk}\Big[\mathrm{Cov}(\varepsilon_{ijk}, q_{ik}) + \mathrm{Cov}(\varepsilon_{ijk}, \eta_{ik})\Big].$$

By Assumption 8(i), $\mathrm{Cov}(q_{ik}, \eta_{ik}) = 0$. By the independence of evaluator signal noise from latent quality (Assumption 5), $\mathrm{Cov}(\varepsilon_{ijk}, q_{ik}) = 0$. By Assumption 8(ii), $\mathrm{Cov}(\varepsilon_{ijk}, \eta_{ik}) = 0$. Hence

$$\mathrm{Cov}(X_{ijk}, s_{ik}) = \omega_{jk} \mathrm{Var}(q_{ik}).$$

Step 2 (Variance of the empirical signal). By Assumption 7 and Assumption 8(i),

$$\mathrm{Var}(s_{ik}) = \mathrm{Var}(q_{ik} + \eta_{ik}) = \mathrm{Var}(q_{ik}) + \mathrm{Var}(\eta_{ik}) + 2\,\mathrm{Cov}(q_{ik}, \eta_{ik}) = \mathrm{Var}(q_{ik}) + \mathrm{Var}(\eta_{ik}).$$

Step 3 (Conclusion). Dividing the results of Steps 1 and 2,

$$\beta_{jk} = \frac{\mathrm{Cov}(X_{ijk}, s_{ik})}{\mathrm{Var}(s_{ik})} = \frac{\omega_{jk} \mathrm{Var}(q_{ik})}{\mathrm{Var}(q_{ik}) + \mathrm{Var}(\eta_{ik})} \equiv \omega_{jk} \lambda_k.$$

Since $\mathrm{Var}(q_{ik})$ and $\mathrm{Var}(\eta_{ik})$ are properties of the venture population and of the measurement procedure and, by Assumption 7, carry no evaluator index, λ_k does not depend on j. Since $\mathrm{Var}(q_{ik}) > 0$ (Assumption 2.2) and $0 < \mathrm{Var}(\eta_{ik}) < \infty$ (Assumption 7), $\lambda_k \in (0,1)$ strictly. □

APPENDIX C: ADDITIONAL TABLES

Criterion	MSE (1)	Var (2)	Bias (3)	MSE/Max(MSE) (4)
Financing	1.45	0.21	1.24	36.2%
Competitive advantage	1.07	0.25	0.82	26.7%
Regulation	1.07	0.20	0.86	26.7%
Skills and connections	0.94	0.20	0.75	23.6%
Business model	0.85	0.22	0.63	21.2%
Projected sales	0.81	0.20	0.61	20.3%
VC interest	0.78	0.15	0.63	19.5%
Feasible solution	0.78	0.22	0.55	19.5%
User pain point	0.77	0.23	0.54	19.1%
Founders speed	0.65	0.17	0.48	16.3%
Technical development status	0.63	0.16	0.47	15.8%
Competition	0.58	0.17	0.41	14.5%
Motivation	0.56	0.25	0.31	13.9%
Customer traction	0.49	0.16	0.33	12.3%
Business development status	0.33	0.14	0.19	8.3%
Average	0.78	0.19	0.59	19.6%

TABLE C.1 Human-AI Grades MSE

The table report for each criterion k the Mean Squared Error (MSE) between human and AI grades (column (1)), which is decomposed into the variance of human grades (column (2)) and the squared bias between the average human grade and the AI grade (column (3)), as shown in the equation below

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\left(X_{ijk}^{H}-X_{ik}^{AI}\right)^2=\frac{1}{N}\sum_{i=1}^{N}\underbrace{\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\left(X_{ijk}^{H}-\bar{X}_{ik}^{H}\right)^2\right]}_{\text{Variance}}+\underbrace{\left[\bar{X}_{ik}^{H}-X_{ik}^{AI}\right]^2}_{\text{Bias}}$$

where N denotes the number of ventures in the sample. Column (4) reports the MSE scaled by the maximum possible MSE.

Dependent Variable:	X_{ijk}			
Model:	(1)	(2)	(3)	(4)
Variables				
$\Delta_{\text{AI-H}}$(Baseline)	0.3697***	0.4597***		
	(0.0328)	(0.0215)		
Human Signal Response	-0.0006	-0.0033	-0.0064	-0.0045
	(0.0076)	(0.0076)	(0.0069)	(0.0054)
$\Delta_{\text{AI-H}}$(Signal Response)	0.0141**	0.0157**	0.0187***	0.0093**
	(0.0058)	(0.0065)	(0.0058)	(0.0038)
Fixed-effects				
Venture Controls	No	Yes	Yes	Yes
Judge-Criterion (1,095)			Yes	
Judge-Venture (2,212)				Yes
Fit statistics				
Human Baseline	2.179	2.002	2.254	2.732
Observations	33,000	33,000	33,000	33,000
R^2	0.05283	0.09920	0.31372	0.38354

TABLE C.2 AI vs. Human Signal Length Response
This tables reports OLS estimates from the regression

$$X_{ijk} = \alpha + D_{j=\text{AI}} + \beta_H s_{i,k} + \beta_{AI}(D_{j=\text{AI}} \times s_{i,k}) + \\ + \gamma_H \mathbf{Z_i} + \gamma_{AI}(D_{j=\text{AI}} \times \mathbf{Z_i}) + + \delta_{j,k} + \lambda_{i,j} + \varepsilon_{ijk}$$

where s_{ik} measures the length of information on venture i about criterion k and it is standardized by criterion; $\mathbf{Z_i}$ is a vector of standardized venture-level controls measured at the time of evaluation, including team size and experience, founder education title, gender and years of experience, number of customer interviewed, capital raised, and incorporation status; δ_{jk} are judge–criterion fixed effects, and λ_{ij} are judge–venture fixed effects."Human Signal Response" is β_H and measures how much human grades change for one standard deviation in information content's length; "$\Delta_{\text{AI-H}}$ (Signal Response)" is the additional AI response β_{AI} and measures whether AI grades change more than human grades for one standard deviation in information content's length; "$\Delta_{\text{AI-H}}$ (Baseline)" is the AI–human level difference in grades α_{AI}. Column (1) includes no controls; column (2) adds venture controls; column (3) adds judge–criterion fixed effects; column (4) adds judge–venture fixed effects. Standard errors clustered at the venture level in parentheses.Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Dependent Variable:			X_{ijk}	
Model:	(1)	(2)	(3)	(4)
Variables				
$\Delta_{\text{AI-H}}$(Baseline)	0.3717***	0.4397***		
	(0.0338)	(0.0246)		
Human Signal Response	0.0514***	0.0473***	0.0512***	-0.0077
	(0.0071)	(0.0068)	(0.0065)	(0.0201)
$\Delta_{\text{AI-H}}$(Signal Response)	0.0278***	0.0282***	0.0243***	0.0125**
	(0.0055)	(0.0052)	(0.0056)	(0.0056)
Fixed-effects				
Venture Controls	No	Yes	Yes	Yes
Judge-Criterion (1,095)			Yes	
Judge-Venture (2,212)				Yes
Fit statistics				
Human Baseline	2.177	2.037	2.296	2.907
Observations	33,000	33,000	33,000	33,000
R^2	0.09742	0.12371	0.33780	0.40279

TABLE C.3 AI vs. Human Application Information Response, controlling for all criteria signals
This tables reports OLS estimates from the regression

$$\begin{aligned} X_{ijk} = \alpha &+ \alpha_{\text{AI}} D_{j=\text{AI}} + \beta_H s_{ik} + \beta_{AI}(D_{j=\text{AI}} \times s_{ik}) \\ &+ \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \left[\beta_H^{k'} s_{ik'} + \beta_{AI}^{k'} (D_{j=\text{AI}} \times s_{ik'}) \right] \\ &+ \gamma_H \mathbf{Z_i} + \gamma_{AI}(D_{j=\text{AI}} \times \mathbf{Z_i}) + \delta_{jk} + \lambda_{ij} + \varepsilon_{ijk}. \end{aligned}$$

where s_{ik} measures the direction of information on venture i about criterion k and it is standardized by criterion; $\mathbf{Z_i}$ is a vector of standardized venture-level controls measured at the time of evaluation, including team size and experience, founder education title, gender and years of experience, number of customer interviewed, capital raised, and incorporation status; δ_{jk} are judge–criterion fixed effects, and λ_{ij} are judge–venture fixed effects."Human Signal Response" is β_H and measures how much human grades on criterion k change for one standard deviation in information content on criterion k, controlling for information on all other criteria $k' \neq k$; "$\Delta_{\text{AI-H}}$ (Signal Response)" is the additional AI response β_{AI} and measures whether AI grades on criterion j change more than human grades for one standard deviation in information content on criterion k, controlling for information on all other criteria $k' \neq k$; "$\Delta_{\text{AI-H}}$ (Baseline)" is the AI–human level difference in grades α_{AI}. Column (1) includes no controls; column (2) adds venture controls; column (3) adds judge–criterion fixed effects; column (4) adds judge–venture fixed effects. Standard errors clustered at the venture level in parentheses.Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Admitted				Rejected			
	Separately		Jointly		Separately		Jointly	
Specification	AI	H	AI	H	AI	H	AI	H
Panel A: Survival								
No controls	0.1467**	0.0904	0.1383**	0.0525	0.1174***	0.1076***	0.0884***	0.0590**
	(0.0561)	(0.0738)	(0.0577)	(0.0728)	(0.0192)	(0.0209)	(0.0241)	(0.0258)
Batch FE	0.1560***	0.0395	0.1607***	-0.0268	0.1155***	0.1041***	0.0880***	0.0563**
	(0.0552)	(0.0843)	(0.0607)	(0.0874)	(0.0195)	(0.0210)	(0.0242)	(0.0257)
Batch FE + Controls	0.1221*	0.0161	0.1273*	-0.0307	0.0838***	0.0639***	0.0713***	0.0351
	(0.0674)	(0.0933)	(0.0743)	(0.1016)	(0.0221)	(0.0239)	(0.0246)	(0.0261)
Batch FE + Controls + Sector FE	0.1262**	0.0256	0.1318*	-0.0284	0.0779***	0.0552**	0.0682***	0.0289
	(0.0603)	(0.0923)	(0.0680)	(0.1023)	(0.0227)	(0.0244)	(0.0250)	(0.0263)
Panel B: Employment $\log(1+L_i)$								
No controls	0.6846***	0.6651***	0.6045***	0.4995***	0.3770***	0.3619***	0.2727***	0.2121***
	(0.1421)	(0.1910)	(0.1377)	(0.1785)	(0.0402)	(0.0461)	(0.0459)	(0.0520)
Batch FE	0.6832***	0.3868*	0.6632***	0.1131	0.3804***	0.3601***	0.2786***	0.2089***
	(0.1258)	(0.2312)	(0.1325)	(0.2095)	(0.0403)	(0.0465)	(0.0458)	(0.0526)
Batch FE + Controls	0.4884***	0.2645	0.4730***	0.0907	0.2885***	0.2526***	0.2317***	0.1591***
	(0.1423)	(0.2375)	(0.1465)	(0.2393)	(0.0441)	(0.0533)	(0.0459)	(0.0556)
Batch FE + Controls + Sector FE	0.4847***	0.2777	0.4677***	0.0858	0.2889***	0.2420***	0.2385***	0.1500***
	(0.1406)	(0.2515)	(0.1531)	(0.2742)	(0.0440)	(0.0551)	(0.0451)	(0.0567)
Panel C: Capital $\mathbf{1}(K_i > 0)$								
No controls	0.1782**	0.2946***	0.1370*	0.2571***	0.0317***	0.0329**	0.0213**	0.0212
	(0.0707)	(0.0774)	(0.0731)	(0.0771)	(0.0094)	(0.0130)	(0.0100)	(0.0144)
Batch FE	0.1735**	0.2026**	0.1486*	0.1413	0.0310***	0.0328**	0.0204**	0.0217
	(0.0750)	(0.0939)	(0.0795)	(0.0948)	(0.0097)	(0.0135)	(0.0102)	(0.0148)
Batch FE + Controls	0.1214	0.1596	0.1005	0.1226	0.0235**	0.0268*	0.0162	0.0203
	(0.0884)	(0.1107)	(0.0929)	(0.1108)	(0.0107)	(0.0162)	(0.0099)	(0.0163)
Batch FE + Controls + Sector FE	0.1234	0.1687	0.0979	0.1286	0.0234**	0.0235	0.0178*	0.0166
	(0.0817)	(0.1107)	(0.0835)	(0.1063)	(0.0114)	(0.0167)	(0.0108)	(0.0168)

TABLE C.4 AI vs. Human Scores and Ventures' Performance: robustness across specifications
This table reports OLS estimates from the regression

$$Y_i = \alpha + \beta_1 AI_i + \beta_2 H_i + \gamma \mathbf{Z}_i + \theta_{b(i)} + \varepsilon_i,$$

where $AI_i = \sum_k X_{ikAI}$ and $H_i = \frac{1}{|J_i|} \sum_{j \in J_i, j \neq AI} \sum_k X_{ikj}$ denote, respectively, the AI and human evaluation scores for venture i, standardized in the pooled sample of ventures. Panel A reports estimates for $Y_i = Survival_i$; Panel B for $Y_i = \log(1+L_i)$, where L_i denotes employment; and Panel C for $Y_i = \mathbf{1}(K_i > 0)$, an indicator for whether the venture raised any capital. Each row reports estimates from a successively richer specification: *No controls* includes neither fixed effects nor venture-level controls; *Batch FE* adds batch fixed effects $\theta_{b(i)}$; *Batch FE + Controls* adds the venture-level controls $\mathbf{Z}_i$ (team size, founder education, customer interviews, capital raised, incorporation status, and business model); *Batch FE + Controls + Sector FE* adds sector fixed effects and corresponds to the baseline specification. Columns report estimates separately for ventures admitted to the accelerator program and for ventures that were not admitted; within each subsample, *Separately* columns include only one evaluator's score at a time, and *Jointly* columns include both scores simultaneously. Heteroskedasticity-robust standard errors are reported in parentheses. Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Dep. Variable: **Survival**	Admitted				Rejected			
	Separately		Jointly		Separately		Jointly	
Criterion	AI	H	AI	H	AI	H	AI	H
Sum of criteria	0.1262**	0.0256	0.1318*	-0.0284	0.0779***	0.0552**	0.0682***	0.0289
	(0.0603)	(0.0923)	(0.0680)	(0.1023)	(0.0227)	(0.0244)	(0.0250)	(0.0263)
Skills and connections	0.0118	-0.0279	0.0120	-0.0280	0.0590***	0.0426*	0.0518**	0.0201
	(0.0649)	(0.0815)	(0.0643)	(0.0821)	(0.0222)	(0.0234)	(0.0242)	(0.0250)
Competition	0.0405*	0.0752	0.0330	0.0643	0.0313*	0.0471**	0.0205	0.0417*
	(0.0225)	(0.0559)	(0.0234)	(0.0580)	(0.0189)	(0.0229)	(0.0199)	(0.0239)
Competitive advantage	0.1335*	0.0132	0.1330*	0.0039	0.0278	0.0441**	0.0215	0.0398*
	(0.0732)	(0.0405)	(0.0748)	(0.0401)	(0.0207)	(0.0221)	(0.0208)	(0.0222)
Regulation	0.0348	0.0497	0.0104	0.0423	0.0365*	0.0253	0.0313	0.0125
	(0.0377)	(0.0437)	(0.0463)	(0.0538)	(0.0205)	(0.0212)	(0.0227)	(0.0233)
Customer traction	0.0208	-0.0124	0.0412	-0.0398	0.0572***	0.0405*	0.0494**	0.0186
	(0.0347)	(0.0537)	(0.0458)	(0.0676)	(0.0220)	(0.0239)	(0.0246)	(0.0266)
Business development status	0.0146	0.0030	0.0147	-0.0004	0.0432**	0.0444*	0.0341*	0.0348
	(0.0455)	(0.0440)	(0.0469)	(0.0456)	(0.0198)	(0.0231)	(0.0206)	(0.0240)
Projected sales	0.0935	-0.0130	0.0981	-0.0290	0.0611***	0.0353	0.0579***	0.0257
	(0.0625)	(0.0531)	(0.0638)	(0.0543)	(0.0207)	(0.0243)	(0.0212)	(0.0249)
Technical development status	0.0429	0.0022	0.0507	-0.0225	0.0523***	0.0465**	0.0404*	0.0295
	(0.0390)	(0.0427)	(0.0470)	(0.0522)	(0.0199)	(0.0232)	(0.0233)	(0.0267)
VC interest	0.0918	0.0045	0.0921	-0.0036	0.0547**	0.0418*	0.0492**	0.0307
	(0.0568)	(0.0558)	(0.0572)	(0.0542)	(0.0212)	(0.0224)	(0.0217)	(0.0225)
Financing	0.0684	0.0775	0.0486	0.0438	0.0158	0.0171	0.0112	0.0137
	(0.0419)	(0.0569)	(0.0474)	(0.0637)	(0.0221)	(0.0202)	(0.0233)	(0.0213)
Founders speed	-0.0041	-0.0656	0.0127	-0.0694	0.0650***	0.0459**	0.0584**	0.0332
	(0.0418)	(0.0510)	(0.0424)	(0.0539)	(0.0216)	(0.0229)	(0.0227)	(0.0235)
User pain point		0.0408		0.0408	-0.0136	0.0057	-0.0146	0.0082
		(0.0744)		(0.0744)	(0.0187)	(0.0236)	(0.0190)	(0.0239)
Feasible solution		0.0581		0.0581	-0.0074	0.0177	-0.0117	0.0204
		(0.0930)		(0.0930)	(0.0211)	(0.0237)	(0.0216)	(0.0240)
Business model		-0.0631		-0.0631	0.0396*	0.0539**	0.0335	0.0478**
		(0.0540)		(0.0540)	(0.0220)	(0.0224)	(0.0216)	(0.0225)
Motivation		-0.0293		-0.0293	-0.0074	0.0084	-0.0083	0.0093
		(0.0750)		(0.0750)	(0.0210)	(0.0223)	(0.0212)	(0.0224)

TABLE C.5 AI, Human Scores and Survival, by Criterion

Notes: This table reports OLS estimates from the regression

$$\text{Survival}_i = \alpha + \beta_1 AI_{ik} + \beta_2 H_{ik} + \gamma \mathbf{Z_i} + \theta_{b(i)} + \varepsilon_i,$$

estimated separately for each criterion k (row 1: the sum across all 15 criteria). AI_{ik} and H_{ik} denote the standardized AI and human grades on criterion k for venture i. In the "Separately" columns, only one evaluator's score is included at a time; in the "Jointly" columns, both AI and human scores are included simultaneously. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects. The Admitted columns report estimates for ventures admitted to the accelerator program; the Rejected columns report estimates for ventures that were not admitted. Heteroskedasticity-robust standard errors are reported in parentheses. Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Dep. Variable: **Employment** $\log(\mathbf{1}+\mathbf{L_i})$	Admitted				Rejected			
	Separately		Jointly		Separately		Jointly	
Criterion	AI	H	AI	H	AI	H	AI	H
Sum of criteria	0.4847***	0.2777	0.4677***	0.0858	0.2889***	0.2420***	0.2385***	0.1500***
	(0.1406)	(0.2515)	(0.1531)	(0.2742)	(0.0440)	(0.0551)	(0.0451)	(0.0567)
Skills and connections	0.1252	-0.0013	0.1253	-0.0025	0.1613***	0.1569***	0.1244***	0.1029*
	(0.1968)	(0.2286)	(0.1984)	(0.2327)	(0.0362)	(0.0474)	(0.0415)	(0.0527)
Competition	0.1664*	0.1375	0.1564*	0.0859	0.3031***	0.2267***	0.2623***	0.1581***
	(0.0850)	(0.1333)	(0.0893)	(0.1420)	(0.0502)	(0.0476)	(0.0512)	(0.0474)
Competitive advantage	0.1916	0.0494	0.1870	0.0363	0.0815**	0.0965**	0.0684*	0.0828*
	(0.2259)	(0.1391)	(0.2339)	(0.1442)	(0.0369)	(0.0454)	(0.0372)	(0.0458)
Regulation	0.1739	0.1281	0.1687	0.0091	0.1238***	0.0661	0.1159**	0.0187
	(0.1126)	(0.1256)	(0.1388)	(0.1541)	(0.0429)	(0.0424)	(0.0474)	(0.0465)
Customer traction	0.3160***	0.0921	0.4079***	-0.1786	0.2274***	0.1688***	0.1926***	0.0834
	(0.1137)	(0.1435)	(0.1463)	(0.1666)	(0.0526)	(0.0547)	(0.0558)	(0.0575)
Business development status	0.2983**	0.1215	0.2807*	0.0563	0.2441***	0.1814***	0.2122***	0.1220**
	(0.1341)	(0.1366)	(0.1440)	(0.1440)	(0.0517)	(0.0509)	(0.0535)	(0.0521)
Projected sales	0.2924**	0.1994	0.2672**	0.1560	0.1843***	0.2396***	0.1573***	0.2134***
	(0.1233)	(0.1604)	(0.1224)	(0.1562)	(0.0379)	(0.0523)	(0.0377)	(0.0524)
Technical development status	0.0833	-0.0188	0.1080	-0.0712	0.1784***	0.1514***	0.1414**	0.0921
	(0.1185)	(0.1400)	(0.1349)	(0.1585)	(0.0479)	(0.0531)	(0.0551)	(0.0606)
VC interest	0.2299**	0.0943	0.2251**	0.0743	0.2029***	0.2140***	0.1710***	0.1754***
	(0.1113)	(0.1987)	(0.1118)	(0.1953)	(0.0393)	(0.0526)	(0.0397)	(0.0530)
Financing	0.0898	-0.0361	0.1544	-0.1432	0.0018	0.0919**	-0.0325	0.1018**
	(0.1104)	(0.1463)	(0.1336)	(0.1703)	(0.0452)	(0.0427)	(0.0464)	(0.0442)
Founders speed	0.1066	0.1607	0.0730	0.1391	0.2433***	0.2002***	0.2128***	0.1538***
	(0.1376)	(0.1662)	(0.1450)	(0.1811)	(0.0409)	(0.0520)	(0.0420)	(0.0515)
User pain point		0.1436		0.1436	0.0284	0.1239***	0.0135	0.1217***
		(0.2201)		(0.2201)	(0.0268)	(0.0446)	(0.0272)	(0.0452)
Feasible solution		0.3061		0.3061	0.0398*	0.1004**	0.0234	0.0950*
		(0.2505)		(0.2505)	(0.0238)	(0.0473)	(0.0251)	(0.0484)
Business model		0.0064		0.0064	0.0673**	0.1935***	0.0433*	0.1856***
		(0.1786)		(0.1786)	(0.0267)	(0.0505)	(0.0248)	(0.0512)
Motivation		-0.1674		-0.1674	0.0262	0.0756	0.0196	0.0734
		(0.1920)		(0.1920)	(0.0298)	(0.0461)	(0.0304)	(0.0465)

TABLE C.6 AI, Human Scores and Employment, by Criterion

Notes: This table reports OLS estimates from the regression

$$\log(1+L_i) = \alpha + \beta_1 AI_{ik} + \beta_2 H_{ik} + \gamma \mathbf{Z_i} + \theta_{b(i)} + \varepsilon_i,$$

estimated separately for each criterion k (row 1: the sum across all 15 criteria). AI_{ik} and H_{ik} denote the standardized AI and human grades on criterion k for venture i. In the "Separately" columns, only one evaluator's score is included at a time; in the "Jointly" columns, both AI and human scores are included simultaneously. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects. The Admitted columns report estimates for ventures admitted to the accelerator program; the Rejected columns report estimates for ventures that were not admitted. Heteroskedasticity-robust standard errors are reported in parentheses. Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Dep. Variable: **Capital** $\mathbf{1}(\mathbf{K_i} \geq \mathbf{0})$	Admitted				Rejected			
	Separately		Jointly		Separately		Jointly	
Criterion	AI	H	AI	H	AI	H	AI	H
Sum of criteria	0.1234	0.1687	0.0979	0.1286	0.0234**	0.0235	0.0178*	0.0166
	(0.0817)	(0.1107)	(0.0835)	(0.1063)	(0.0114)	(0.0167)	(0.0108)	(0.0168)
Skills and connections	-0.0710	0.0738	-0.0717	0.0745	0.0208**	0.0133	0.0190*	0.0051
	(0.0812)	(0.1196)	(0.0807)	(0.1188)	(0.0087)	(0.0156)	(0.0098)	(0.0171)
Competition	0.0517	-0.0112	0.0551	-0.0294	0.0311*	0.0243*	0.0267	0.0174
	(0.0490)	(0.0808)	(0.0494)	(0.0775)	(0.0162)	(0.0137)	(0.0167)	(0.0140)
Competitive advantage	-0.0686	0.0426	-0.0746	0.0478	0.0159**	0.0257*	0.0122	0.0232
	(0.1017)	(0.0676)	(0.1076)	(0.0693)	(0.0073)	(0.0146)	(0.0074)	(0.0147)
Regulation	0.0767	-0.0301	0.1584**	-0.1419*	-0.0062	0.0182	-0.0167	0.0250**
	(0.0507)	(0.0678)	(0.0608)	(0.0790)	(0.0139)	(0.0124)	(0.0141)	(0.0123)
Customer traction	0.0739	0.0389	0.0819	-0.0155	0.0091	0.0251	-0.0016	0.0258
	(0.0692)	(0.0732)	(0.0801)	(0.0837)	(0.0132)	(0.0171)	(0.0141)	(0.0186)
Business development status	-0.0046	0.0913	-0.0358	0.0996	0.0087	0.0012	0.0090	-0.0013
	(0.0672)	(0.0631)	(0.0672)	(0.0642)	(0.0147)	(0.0156)	(0.0152)	(0.0160)
Projected sales	-0.0085	0.1872***	-0.0398	0.1937***	0.0349***	0.0233*	0.0326***	0.0179
	(0.0730)	(0.0637)	(0.0825)	(0.0683)	(0.0103)	(0.0135)	(0.0101)	(0.0133)
Technical development status	0.0392	-0.0940	0.0864	-0.1360*	-0.0040	0.0128	-0.0109	0.0174
	(0.0593)	(0.0636)	(0.0642)	(0.0712)	(0.0109)	(0.0143)	(0.0126)	(0.0163)
VC interest	-0.0117	0.2915***	-0.0307	0.2943***	0.0329***	0.0246	0.0296***	0.0179
	(0.0728)	(0.0713)	(0.0713)	(0.0729)	(0.0114)	(0.0154)	(0.0111)	(0.0151)
Financing	0.1296**	0.0094	0.1825**	-0.1172	-0.0159	-0.0018	-0.0170	0.0034
	(0.0546)	(0.0620)	(0.0716)	(0.0777)	(0.0126)	(0.0120)	(0.0123)	(0.0117)
Founders speed	-0.0534	0.0525	-0.0712	0.0736	0.0348***	0.0222	0.0317***	0.0153
	(0.0757)	(0.0796)	(0.0769)	(0.0809)	(0.0124)	(0.0163)	(0.0121)	(0.0162)
User pain point		0.1899**		0.1899**	0.0049	0.0085	0.0039	0.0079
		(0.0944)		(0.0944)	(0.0032)	(0.0136)	(0.0036)	(0.0139)
Feasible solution		0.1116		0.1116	0.0060	0.0046	0.0052	0.0034
		(0.0857)		(0.0857)	(0.0039)	(0.0130)	(0.0044)	(0.0134)
Business model		-0.0341		-0.0341	0.0108**	0.0088	0.0099**	0.0070
		(0.0798)		(0.0798)	(0.0048)	(0.0136)	(0.0050)	(0.0138)
Motivation		0.0990		0.0990	0.0051	0.0124	0.0040	0.0119
		(0.0863)		(0.0863)	(0.0033)	(0.0130)	(0.0033)	(0.0131)

TABLE C.7 AI, Human Scores and Capital, by Criterion

Notes: This table reports OLS estimates from the regression

$$\mathbf{1}(K_i > 0) = \alpha + \beta_1 AI_{ik} + \beta_2 H_{ik} + \gamma \mathbf{Z_i} + \theta_{b(i)} + \varepsilon_i,$$

estimated separately for each criterion k (row 1: the sum across all 15 criteria). AI_{ik} and H_{ik} denote the standardized AI and human grades on criterion k for venture i. In the "Separately" columns, only one evaluator's score is included at a time; in the "Jointly" columns, both AI and human scores are included simultaneously. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects. The Admitted columns report estimates for ventures admitted to the accelerator program; the Rejected columns report estimates for ventures that were not admitted. Heteroskedasticity-robust standard errors are reported in parentheses. Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Specification	Description
Role-Based (Baseline)	Casts the AI in the role of a reviewer at HEC Incubateur and instructs it to base its assessment exclusively on the application and on information available as of the application's submission date.
Zero-Shot	Replaces the reviewer persona with a generic statement of the scoring task, isolating the contribution of the role framing itself.
Few-Shot	Provides three worked examples illustrating how the scoring rubric applies to hypothetical application excerpts on three separate criteria.
Chain of Thought	Instructs the model to work through five explicit analytical steps, covering team capacity, problem-solution fit, competitive positioning, and commercial and financial traction, before assigning scores.
Self-Consistency	Instructs the model to perform three internal evaluation passes, an evidentiary pass restricted to verifiable facts, a comparative pass benchmarking the venture against typical start-ups, and a skeptical pass searching for gaps or overstated claims, and to set the final score to the majority outcome across the three passes.
Reflection on Search Trees	Instructs the model to weigh, criterion by criterion, the strongest evidence supporting a high score against the strongest evidence supporting a low score before committing to a final assessment.
Application Question Order	Retains the baseline instructions in full and instead reorders the sequence in which the venture's application answers are presented to the model, replacing the original questionnaire order with a semi-random block order that groups questions addressing a common topic, team composition, market sizing, or financing, for instance, into contiguous sections.

TABLE C.8 AI Prompt Specifications

Model	Release	Size	Training cutoff	Architecture / weights
Llama-3.3-70B-Instruct (Baseline)	Dec 2024	70B	Dec 2023	Dense transformer; open
DeepSeek R1	Jan 2025	70B	∼Jul 2024	Sparse MoE; open
Qwen2.5-7B-Instruct	Sep 2024	7B	∼Jun 2024	Dense transformer; open
Gemma-3n-E4B	Jun 2025	4B	∼mid-2024	Nested sub-models; open
Claude Sonnet 4.6	Feb 2026	Undisclosed	Aug 2025	Undisclosed; closed

TABLE C.9 AI Model Specifications

Batch	N Ventures	Survival Mean	FTE Mean	FTE $P75$	€Mln Funding $P>0$	€Mln Funding Mean$\mid>0$	€Mln Funding $P75\mid>0$
2021 Fall	62	60%	10.5	11	29%	12	2.2
2021 Summer	49	59%	10.1	13	41%	4	0.4
2022 Spring	60	68%	8.2	11	42%	1.4	1
2022 Summer	96	59%	13	11	24%	3.8	1.8
2022 Winter	66	61%	17.2	18.8	35%	5.1	1
2023 Spring	72	53%	6.2	7	18%	0.6	0.7
2023 Summer	66	70%	4.1	6	29%	0.6	0.6
2023 Winter	54	70%	5	5.8	22%	0.7	0.9
2024 Winter	54	63%	5.6	8.8	31%	0.4	0
All	579	62%	9	10	29%	3.4	1

TABLE C.10 Ventures Survival Rates, FTE and Funding, medium term
The table presents summary statistics for all ventures evaluated across all batches in the dataset. Column (2) reports the number of ventures evaluated; column (3) shows the survival rate, i.e., the percentage of ventures in each batch still active as of Spring 2026. Columns (4)-(5) shows the average and the 75th percentile of FTEs among surviving ventures. Columns (6)-(8) provide metrics on post application funding in millions of euros: the probability of receiving any funding post-application, the mean funding amount if received, and the funding amount for the top 25% by post application funding in the batch.

	Admitted			Rejected		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Survival						
AI Score$_i$	0.1483		0.1562	0.0270		0.0094
	(0.0938)		(0.1000)	(0.0310)		(0.0330)
AI Score$_i \times \mathbf{1}(b(i) < 2023)$	-0.0468		-0.1090	0.1081**		0.1241**
	(0.1165)		(0.1391)	(0.0464)		(0.0512)
H Score$_i$		-0.0603	-0.0875		0.0529*	0.0507
		(0.1168)	(0.1213)		(0.0318)	(0.0345)
H Score$_i \times \mathbf{1}(b(i) < 2023)$		0.3111	0.2723		0.0102	-0.0450
		(0.2244)	(0.2660)		(0.0489)	(0.0526)
R^2	0.55133	0.52567	0.56293	0.23416	0.21009	0.23785
Panel B: Employment $\log(1+L_i)$						
AI Score$_i$	0.4803***		0.4778***	0.2189***		0.1711***
	(0.1603)		(0.1632)	(0.0533)		(0.0554)
AI Score$_i \times \mathbf{1}(b(i) < 2023)$	-0.0232		-0.0272	0.1264		0.1168
	(0.2424)		(0.3434)	(0.0921)		(0.0950)
H Score$_i$		0.1175	0.0235		0.2039***	0.1415**
		(0.2580)	(0.2454)		(0.0671)	(0.0718)
H Score$_i \times \mathbf{1}(b(i) < 2023)$		0.4745	-0.0071		0.0844	0.0246
		(0.5604)	(0.7236)		(0.1113)	(0.1140)
R^2	0.71620	0.67538	0.71625	0.31272	0.29020	0.32564
Panel C: Capital $\mathbf{1}(K_i \geq 0)$						
AI Score$_i$	0.1917*		0.1799	0.0083		0.0059
	(0.1113)		(0.1173)	(0.0134)		(0.0154)
AI Score$_i \times \mathbf{1}(b(i) < 2023)$	-0.2405		-0.4993**	0.0268		0.0172
	(0.1911)		(0.1902)	(0.0224)		(0.0210)
H Score$_i$		0.0900	0.0717		0.0097	0.0077
		(0.1180)	(0.1130)		(0.0188)	(0.0211)
H Score$_i \times \mathbf{1}(b(i) < 2023)$		0.3693	0.7671**		0.0342	0.0263
		(0.3184)	(0.3343)		(0.0333)	(0.0335)
R^2	0.51354	0.51116	0.56117	0.14258	0.14449	0.14699
Model Specification						
Batch FE	Yes	Yes	Yes	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes	Yes	Yes	Yes
Venture controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	93	93	93	473	473	473

TABLE C.11 AI vs. Human Scores and Venture performance: Split by early Batches
This table reports OLS estimates from the regression

$$Y_i = \alpha + \beta_1 AI_i + \beta_2 (AI_i \times Early_i) + \beta_3 H_i + \beta_4 (H_i \times Early_i) + (\gamma \mathbf{Z}_i) \times (1 + Early_i) + \theta_{b(i)} + \varepsilon_i,$$

where $AI_i = \sum_k X_{ikAI}$ and $H_i = \frac{1}{|J_i|} \sum_{j \in J_i, j \neq AI} \sum_k X_{ikj}$ denote, respectively, the AI and human evaluation scores for venture i, standardized in the pooled sample of ventures, and $Early_i = \mathbf{1}(b(i) < 2023)$ indicates a venture evaluated in a batch before 2023. $\mathbf{Z}_i$ is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised prior to application, incorporation status, and business model, each also interacted with $Early_i$. $\theta_{b(i)}$ $\delta_{s(i)}$ are batch fixed effects. Columns (1)–(3) report estimates for ventures admitted to the accelerator program, while columns (4)–(6) report estimates for ventures that were not admitted. In columns (1) and (4), only AI scores are included; in columns (2) and (5), only human scores are included; in columns (3) and (6), both AI and human scores are included jointly. Heteroskedasticity-robust standard errors are reported in parentheses. Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX D: ADDITIONAL FIGURES

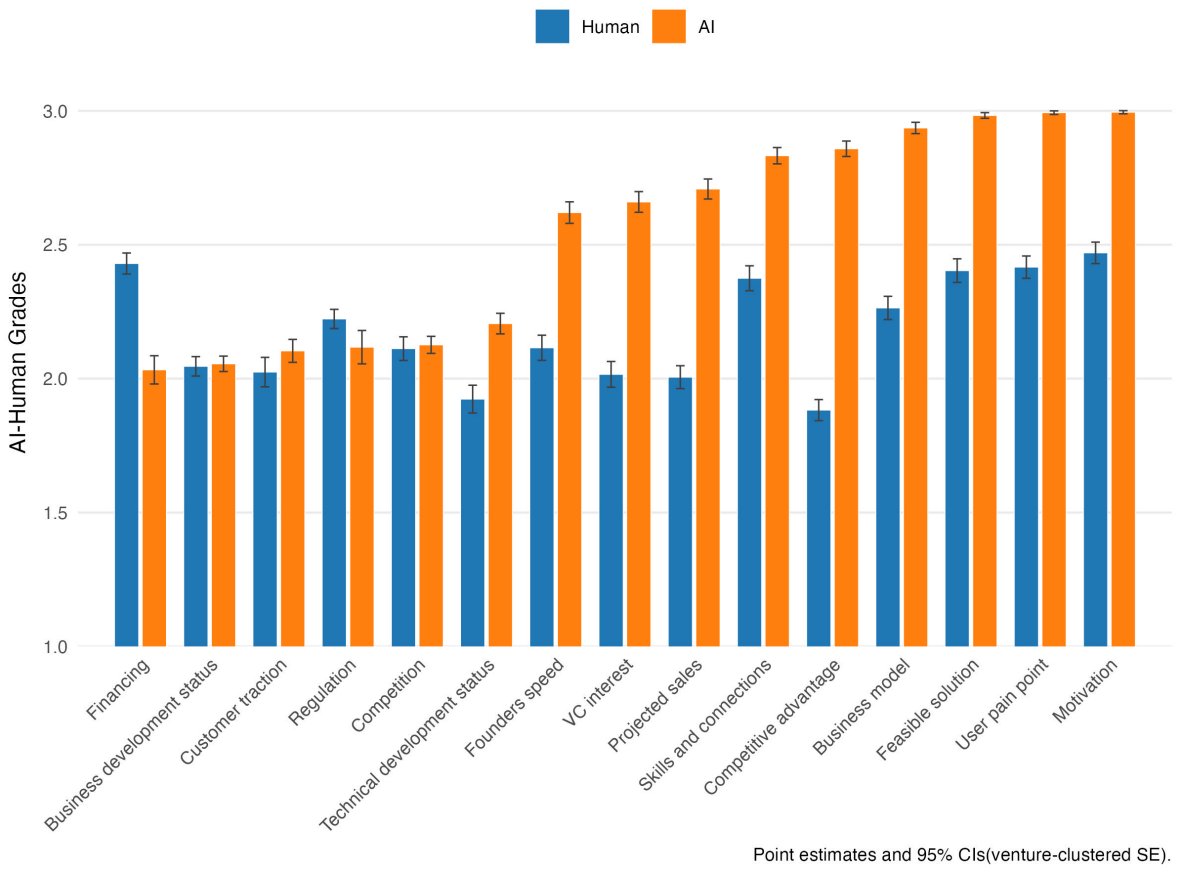

FIGURE D.1 AI vs. Human Grades by Criterion.
For each criterion $k = 1, \ldots, 15$, the figure reports the estimated average grade and 95% confidence interval for human judges (blue) and for the AI (green). Estimates are obtained from the regression

$$X_{ijk} = \alpha + \sum_{k=1}^{15} \beta^k D_k + \beta_{\text{AI}} D_{j=\text{AI}} + \sum_{k=1}^{15} \beta_{\text{AI}}^k D_k D_{j=\text{AI}} + \varepsilon_{ijk},$$

where X_{ijk} is the grade of venture i on criterion k by judge j, D_k is a dummy for criterion k, $D_{j=\text{AI}}$ is an indicator for the AI. AI grades to criterion k of venture i are appended to the venture $\times$ criterion $\times$ judge dataset so that the AI is treated as an additional judge. The reference criterion is $k =$ "Founders speed": for this criterion, the human bar corresponds to the intercept, while the AI bar corresponds to the intercept plus β_{AI}. For the other criteria, the human bar adds the corresponding β^k term, and the AI bar additionally includes the interaction coefficient β_{AI}^k. Standard errors are clustered at the venture level.

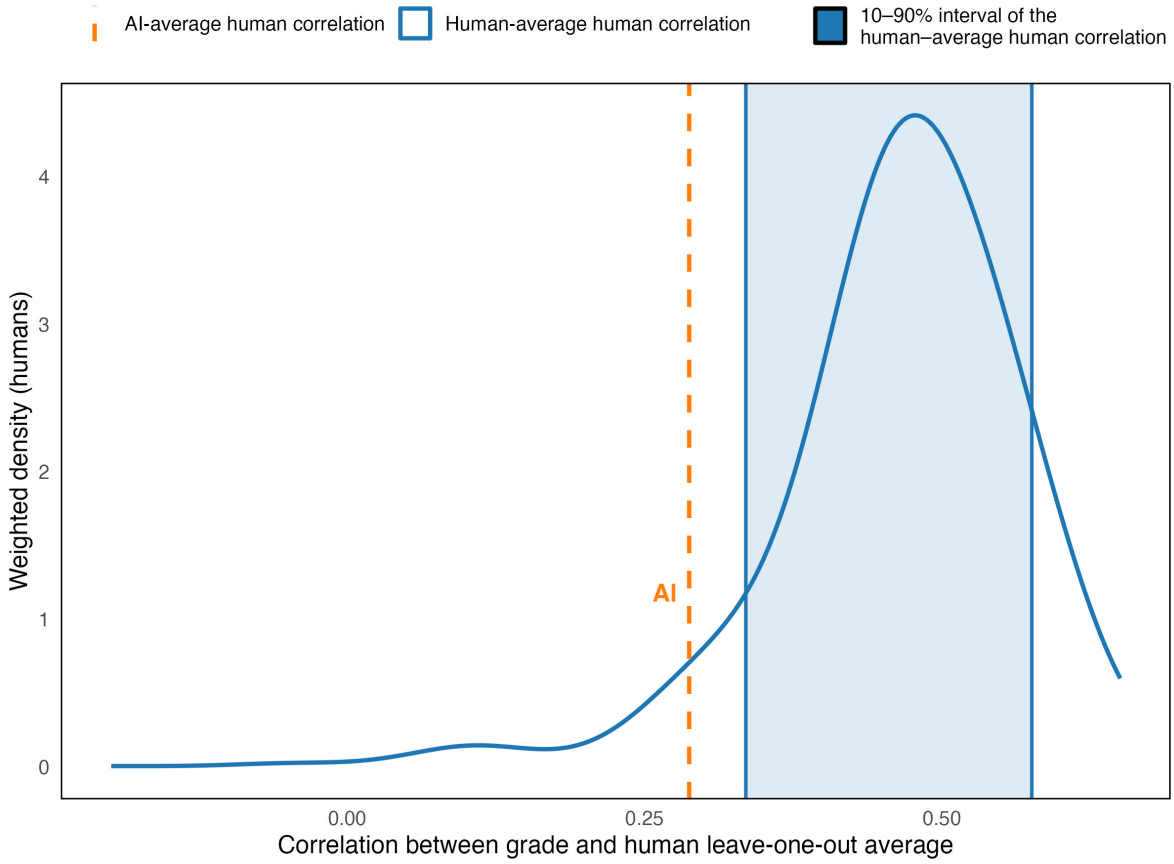

FIGURE D.2 AI vs human correlation with other human judges
The graph shows the distribution of the correlation between each individual judge's grades and the average grade of the other judges for each venture–criterion pair. Each judge's correlation is weighted by the number of evaluations they made. The vertical orange line reports the correlation between the AI's grades and the average of human grades within each venture–criterion pair, where one human is randomly excluded to make it comparable to the human leave-one-out measure. The shaded area represents the 10–90% interval of the human–human correlation distribution.

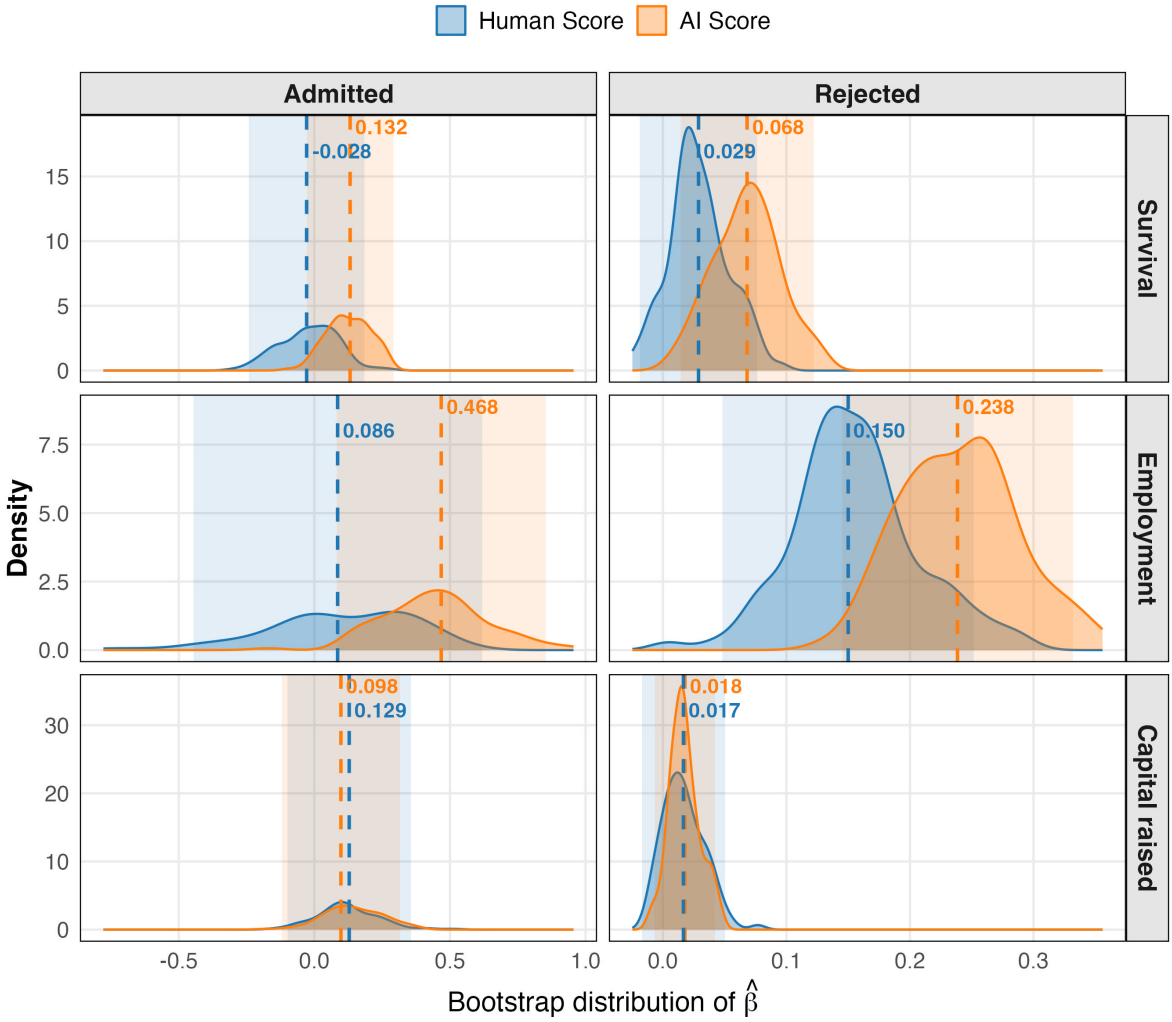

FIGURE D.3 Human vs AI Scores and Venture Performance, Bootstrap estimates
This figure reports the nonparametric bootstrap distribution of the coefficients on the standardized AI and human evaluation scores in the joint specification

$$Y_i = \alpha + \beta_1 AI_i + \beta_2 H_i + \gamma \mathbf{Z_i} + \theta_{b(i)} + \varepsilon_i.$$

estimated separately for ventures admitted and rejected from the HEC Incubateur. The first row reports estimates for $Y_i = Survival_i$; the second for for $Y_i = \log(1 + L_i)$, where L_i denotes employment; the third for $Y_i = \mathbf{1}(K_i > 0)$, an indicator for whether the venture raised any capital. $AI_i = \sum_k X_{ikAI}$ and $H_i = \frac{1}{|J_i|} \sum_{j \in J_i, j \neq \text{AI}} \sum_k X_{ikj}$ denote, respectively, the AI and human evaluation scores for venture i, and they are standardized in the pooled sample of ventures. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects. Each panel plots the kernel density of the bootstrap distribution of $\hat{\beta}_1$ (orange) and $\hat{\beta}_2$ (blue) across $B = 100$ stratified nonparametric bootstrap replications: ventures are resampled with replacement separately within the admitted and rejected subsamples, and the joint specification is re-estimated on every replicate. Dashed vertical lines mark the point estimate from the original, non-resampled sample; shaded bands span two bootstrap standard deviations around that estimate.

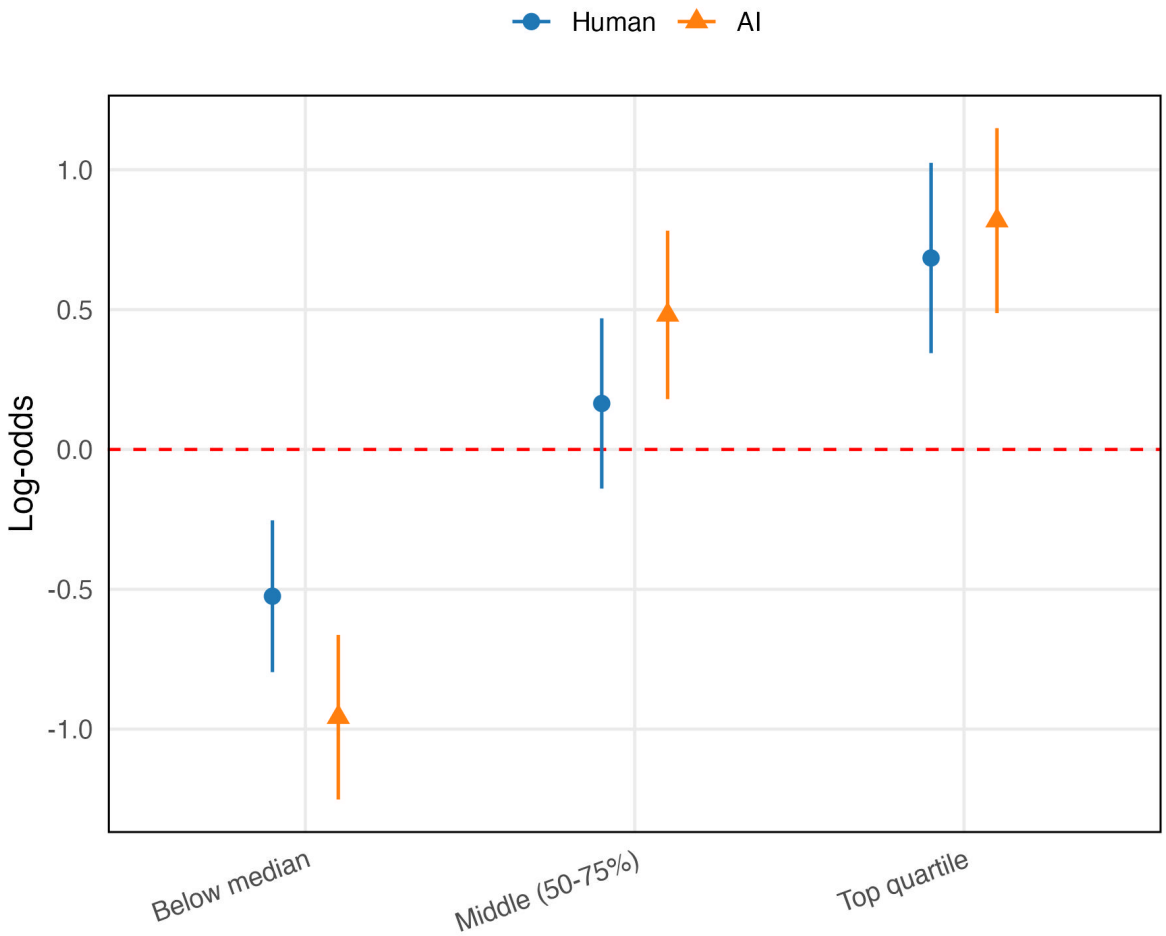

FIGURE D.4 Human vs AI Scores and Employment along distribution
This figure reports coefficients from the logit model

$$\Pr(Y_i \in S_m) = \Lambda\left(\alpha_m + \beta_{1m} AI_i + \beta_{2m} H_i + \gamma_m \mathbf{Z}_i + \theta_{b(i)}\right), \qquad m \in \{\text{below median, middle, top quartile}\},$$

estimated separately for each segment m of the log-employment distribution, where $\Lambda(\cdot)$ is the logistic cumulative distribution function, pooling admitted and rejected ventures. $AI_i = \sum_k X_{ikAI}$ and $H_i = \frac{1}{|J_i|} \sum_{j \in J_i, j \neq \text{AI}} \sum_k X_{ikj}$ denote, respectively, the AI and human evaluation scores for venture i, and they are standardized in the pooled sample of ventures. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects; the specification additionally controls for admission status. Blue circles report β_{2m}, the coefficient on the human score; orange triangles report β_{1m}, the coefficient on the AI score. Vertical lines report 95% confidence intervals. The dashed line marks a log-odds of zero.

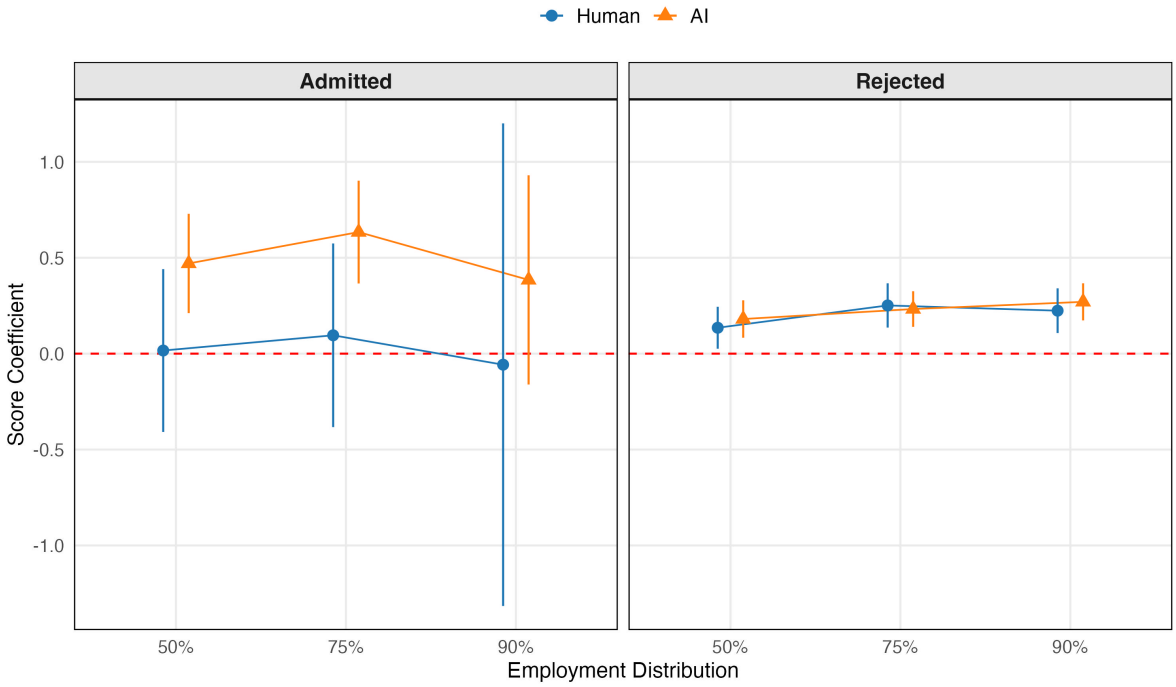

FIGURE D.5 Human vs AI Scores and Employment, by quantile
This figure reports estimates from quantile regressions of log employment on AI and human evaluation scores, estimated at the, 50th, 75th, and 90th percentiles, separately among admitted and rejected ventures:

$$\log(1+L_i) = \alpha + \beta_1 AI_i + \beta_2 H_i + \gamma \mathbf{Z_i} + \theta_{b(i)} + \varepsilon_i.$$

$AI_i = \sum_k X_{ik\text{AI}}$ and $H_i = \frac{1}{|J_i|} \sum_{j \in J_i,\, j \neq \text{AI}} \sum_k X_{ikj}$ denote, respectively, the AI and average human evaluation scores for venture i, where J_i is the set of human judges evaluating venture i, and they are standardized in the sample of ventures. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size , number of customer interviews, capital raised, IT sector dummy, and incorporation status; $\theta_{b(i)}$ are batch fixed effects. Each panel plots the point estimates and 95% confidence intervals (heteroskedasticity-robust standard errors) for β_1 (AI, orange) and β_2 (human, blue) at each quantile of the log-employment distribution. The left panel reports estimates for ventures admitted to the HEC Incubateur; the right panel reports estimates for rejected ventures.

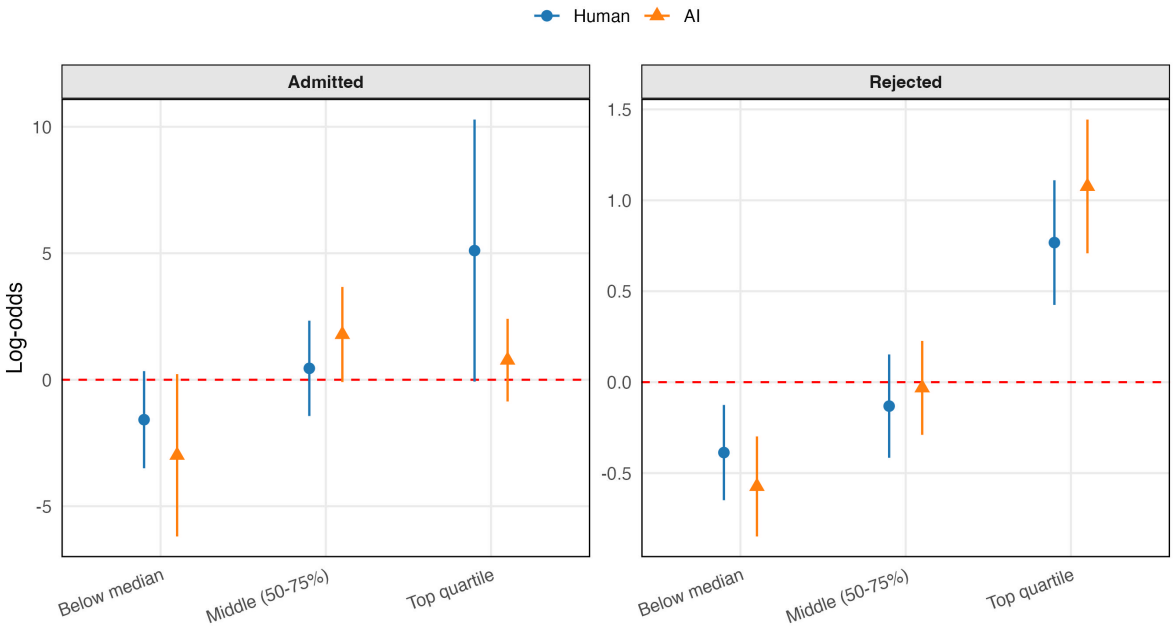

FIGURE D.6 Human vs AI Scores and Employment along distribution
This figure reports coefficients from the logit model

$$\Pr\left(Y_i \in S_m\right) = \Lambda\left(\alpha_m + \beta_{1m} AI_i + \beta_{2m} H_i + \gamma_m \mathbf{Z}_i + \theta_{b(i)}\right), \qquad m \in \{\text{below median, middle, top quartile}\},$$

estimated separately for each segment m of the log-employment distribution, where $\Lambda(\cdot)$ is the logistic cumulative distribution function, separately among admitted and rejected ventures. $AI_i = \sum_k X_{ikAI}$ and $H_i = \frac{1}{|J_i|}\sum_{j \in J_i, j \neq \text{AI}} \sum_k X_{ikj}$ denote, respectively, the AI and human evaluation scores for venture i, and they are standardized in the pooled sample of ventures. $\mathbf{Z_i}$ is a vector of venture-level controls measured at the time of evaluation, including team size, founder education, customer interviews, capital raised, incorporation status, sector, and business model; $\theta_{b(i)}$ are batch fixed effects. Blue circles report β_{2m}, the coefficient on the human score; orange triangles report β_{1m}, the coefficient on the AI score. The left panel reports estimates for ventures admitted to the HEC Incubateur; the right panel reports estimates for rejected ventures. Vertical lines report 95% confidence intervals. The dashed line marks a log-odds of zero.

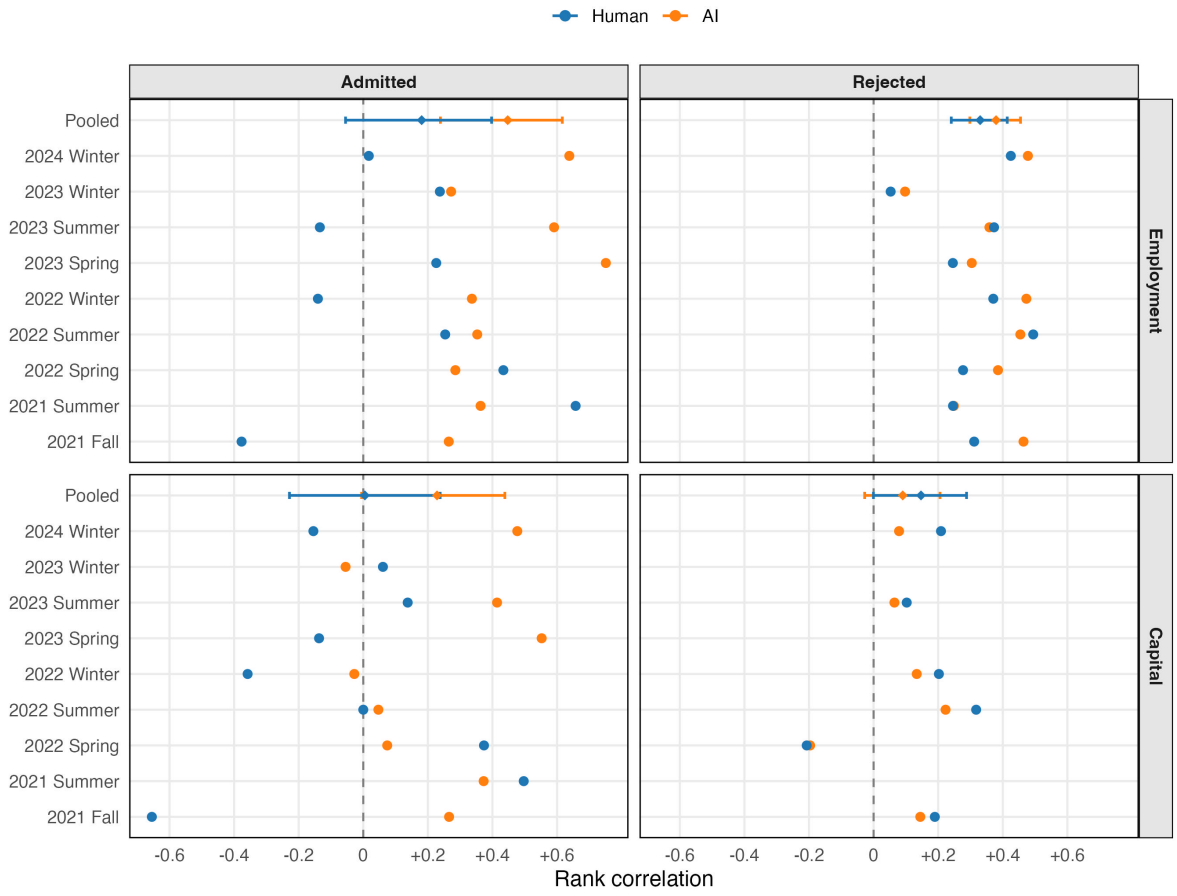

FIGURE D.7 Evaluators and performance Ranking, by Admission Status
This figure reports, for each batch, the correlation between each evaluator's ranking of ventures by the score they assigned, defined as the sum of criterion grades, and the realized outcome, for employment (top row) and capital raised (bottom row), separately for admitted (left pannel) and rejected (right panel) within each batch. Blue points report the correlation with human ranking; orange points report the correlation with AI ranking. The "Pooled" row in each panel aggregates the batch-level correlations via Fisher-z-transformed random-effects (REML) meta-analysis and reports the back-transformed pooled correlation together with its 95% confidence interval. Batch-level correlations are point estimates only, no confidence interval is shown for them individually, since each cell contains too few ventures to estimate one with any precision.

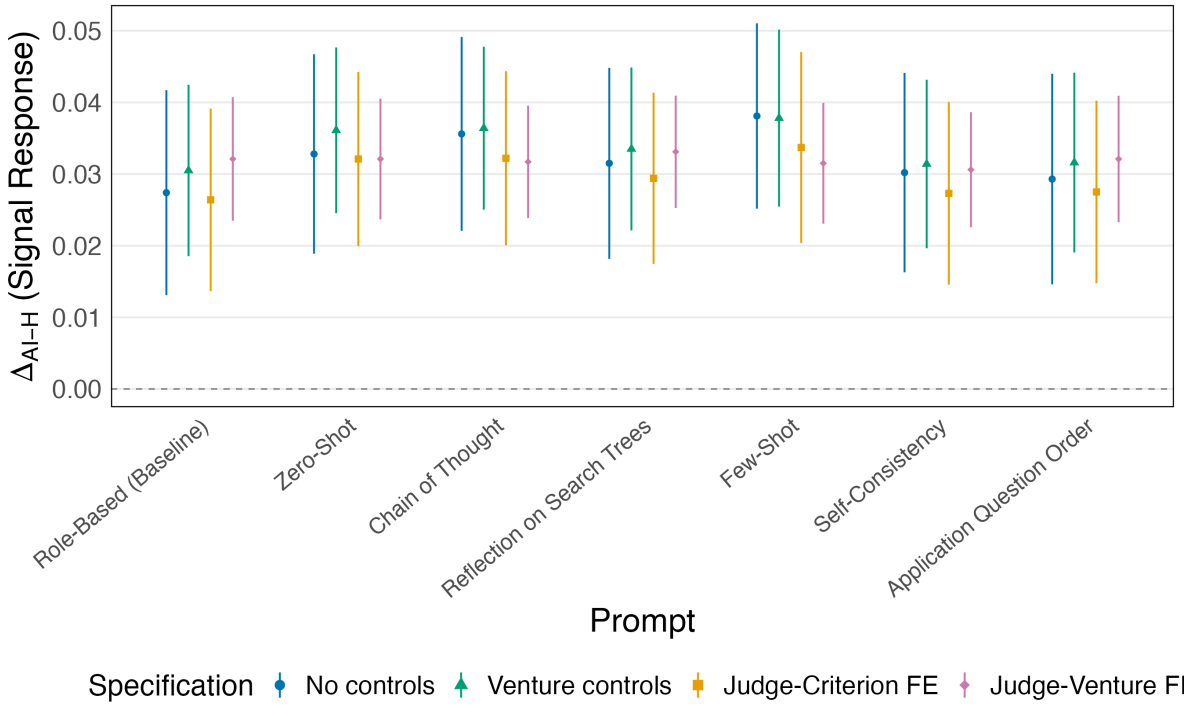

FIGURE D.8 AI vs. Human signal response, by prompt
This figure reports estimates of $\Delta_{\text{AI-H}}(\text{Signal Response})$, the additional responsiveness of AI grades to criterion-relevant application content relative to human judges, from the regression

$$X_{ijk} = \alpha + \alpha_{AI} D_{j=AI} + \beta_H s_{ik} + \beta_{AI}(D_{j=AI} \times s_{ik}) + \gamma_H \mathbf{Z_i} + \gamma_{AI}(D_{j=AI} \times \mathbf{Z_i}) + \delta_{jk} + \lambda_{ij} + \varepsilon_{ijk},$$

estimated separately under each of the seven prompt specifications used to generate the AI grades. s_{ik} measures the direction of information on venture i about criterion k and is standardized by criterion; $\mathbf{Z_i}$ is a vector of standardized venture-level controls measured at the time of evaluation, including team size and experience, founder education, gender and years of experience, number of customers interviewed, capital raised, and incorporation status; δ_{jk} are judge-criterion fixed effects, and λ_{ij} are judge-venture fixed effects. β_H is the human signal response and measures how much human grades change for one standard deviation in information content; $\Delta_{\text{AI-H}}(\text{Signal Response})$ is the additional AI response β_{AI} that measures whether AI grades change more than human grades for one standard deviation in information content, the coefficient of interest plotted on the vertical axis. For each prompt, the graph plots the estimated coefficient $\Delta_{\text{AI-H}}(\text{Signal Response})$ under an increasing set of controls: no controls, venture controls, venture controls plus judge-by-criterion fixed effects, and venture controls plus judge-by-venture fixed effects. Vertical lines report 95% confidence intervals from standard errors clustered at the judge and venture level. Vertical lines report 95% confidence intervals from standard errors clustered at the judge and venture level.

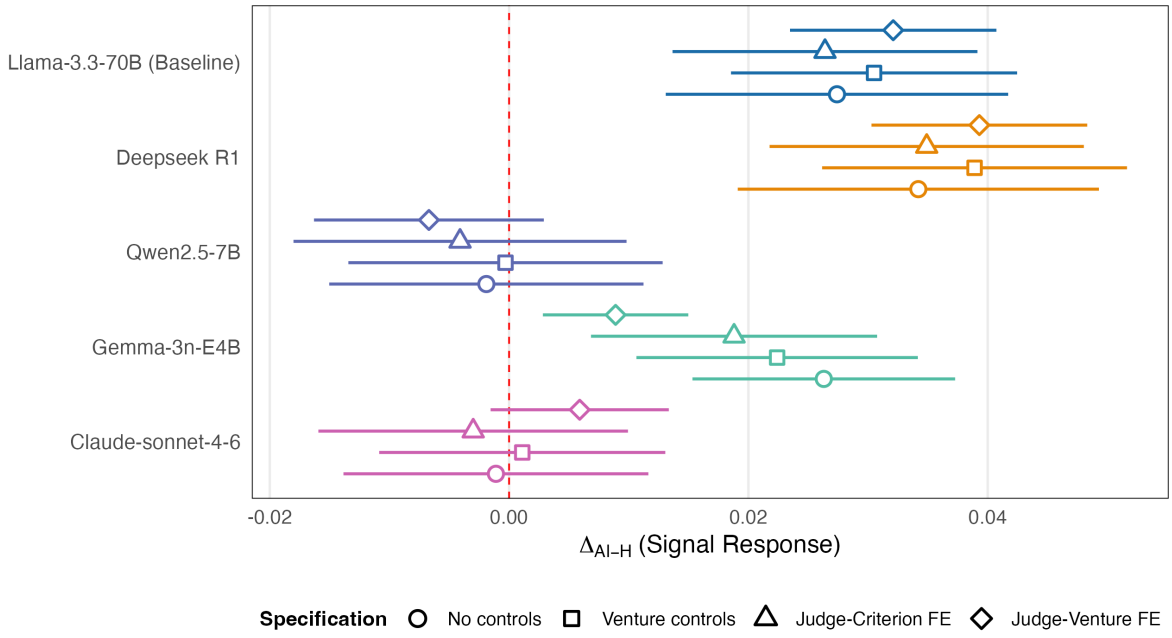

FIGURE D.9 AI vs. Human signal response, by AI model
This figure reports estimates of $\Delta_{\text{AI-H}}(\text{Signal Response})$, the additional responsiveness of AI grades to criterion-relevant application content relative to human judges, from the regression

$$X_{ijk} = \alpha + \alpha_{AI} D_{j=AI} + \beta_H s_{ik} + \beta_{AI}(D_{j=AI} \times s_{ik}) + \gamma_H \mathbf{Z_i} + \gamma_{AI}(D_{j=AI} \times \mathbf{Z_i}) + \delta_{jk} + \lambda_{ij} + \varepsilon_{ijk},$$

estimated separately under each of the five different models used to generate the AI grades. s_{ik} measures the direction of information on venture i about criterion k and is standardized by criterion; $\mathbf{Z_i}$ is a vector of standardized venture-level controls measured at the time of evaluation, including team size and experience, founder education, gender and years of experience, number of customers interviewed, capital raised, and incorporation status; δ_{jk} are judge-criterion fixed effects, and λ_{ij} are judge-venture fixed effects. β_H is the human signal response and measures how much human grades change for one standard deviation in information content; $\Delta_{\text{AI-H}}(\text{Signal Response})$ is the additional AI response β_{AI} that measures whether AI grades change more than human grades for one standard deviation in information content, the coefficient of interest plotted. For each AI model, the graph plots the estimated coefficient $\Delta_{\text{AI-H}}(\text{Signal Response})$ under an increasing set of controls: no controls, venture controls, venture controls plus judge-by-criterion fixed effects, and venture controls plus judge-by-venture fixed effects. Vertical lines report 95% confidence intervals from standard errors clustered at the judge and venture level. Vertical lines report 95% confidence intervals from standard errors clustered at the judge and venture level.

APPENDIX E: APPLICATION FORMS

Below I report the application form filled out by applying companies.

1. Select the batch to which you apply
    - Normally there should only be one batch available.
2. Which Program do you wish to apply to?
    - one with desks at Station F (300€ excl. tax / Desk / Month).
    - one fully remote & online (300€ excl. tax / Startup / Month).
3. Please indicate the first name of the responsible for your application;
4. Please indicate the last name of the person responsible for your application;
5. Please provide the email address of the person responsible for your application;
6. Please add your mobile number
7. Please add a link to your LinkedIn profile
8. What's your job / role inside the company?
9. What's your link with HEC Paris?
10. Which of the following applies to at least one of your co-founders?
    - I am a HEC Alumni
    - I am a HEC Student
    - One of my co-founder is a HEC Student or Alumni
    - I have a HEC Certificate
    - I am in the process of obtaining a HEC Certificate like a MooC
    - I am affiliated with a partner from Incubateur HEC
    - No link with HEC
11. What school /university did you attend?
12. What degree did you obtain at that institution?
13. What year did you graduate?
14. Do you have another degree? If so, please indicate the university and degree name
15. What is your company name? *
16. How many co-founders constitute your team? *

17. What is the gender mix of the cofounders? *
18. Please describe your company in one sentence. Please do not type more than 240 characters.
19. Please add your logo (in a png format)
20. Please add a link to your website
21. Have you already incorporated your company ? *
22. How much time will you dedicate to the project on the coming year? (in % of your working time).This question concerns all the founders.
    - 100%
    - 80% – 99%
    - 75% – 79%
    - 50% – 74%
    - $<$ 50%
23. How many people work for your startup?
24. Why are your team and you the most qualified to make the project happen? (describe your team: name, role, contract, experience)
25. What is your business model?
    - Agency / Training center: you sell your time and your expertise (revenues = contract)
    - Subscription: you sell a recurring access to a product/service (revenues = MRR)
    - Marketplace: you connect buyers to sellers (revenues = fees on each transaction)
    - (e)Commerce: you buy or create products/services and resell it (revenues = margin)
    - Mobile App: you offer captive environments where the users must buy add ons in order to reach their goal (revenues = in-app purchases)
    - Media/Social networks: you sell audience attention to advertisers (revenues = ads / affiliation / brand content)
26. What is the maturity stage of your solution?
    - Idea: an interactive demo
    - Prototype
    - Minimum Viable Product
    - Full working Product
27. How long is your startup runway?
28. How long can keep working on your project before cash becomes an issue?

- I don't know
- Less than 3 months
- 3 - 6 months
- 6 - 12 months
- More than 12 months

29. How many workstations will you need when you join the incubator?

30. How did you hear about Incubateur HEC Paris?

    - A startup from the Incubateur HEC Paris
    - A startup from StationF
    - An event at StationF or StationF's website
    - HEC Paris website
    - A VC Fund
    - Word of mouth
    - An expert
    - An incubee
    - Social networks / Press
    - Other
    - Newsletter of the Incubateur HEC Paris
    - Newsletter of the Innovation & Entrepreneurship Center

31. Please provide us the name of a startup or a person that would be ready to recommend your application

32. Please, confirm you have read our data processing agreement : https://tinyurl.com/DPA-IncubateurHEC

33. Do you agree with our values? Incubateur HEC Paris is a community of entrepreuneurs, and that community is only as strong as the engagement of its members. We have always believed that each startup is different, and that each is deserving of different resources at different points in their journey. But that diversity makes ever more important the shared commitment that we ask of all our incubees. Therefore, all entrepreneurs who join us pledge to:

    - Be available in Paris at the start of the program for the two-day onboarding, alongside the rest of their batch.
    - Dedicate a 45-minute recurring slot every two weeks on Tuesday or Wednesday morning
    - Get involved in the collaborative growth of the Incubateur by sharing their own network and experience with other incubees.

34. Please provide a link to a 2 minute video of you presenting your startup
35. Please provide a link to a 2 minute video presenting your team
36. Please upload your investor deck (pdf only)
37. How long ago did you start to work full-time on your project?
38. How many people work for your startup?
39. Why are your team and you the most qualified to make the project happen? (describe your team: name, role, contract, experience)
40. Among your teammates, how long is the longest professional experience in the sector/market that you are targeting?
41. Who are your current mentors regarding your company? (different from advisory board members)
42. Do you have an advisory board? If so, does it include a Director of a leading company of your target sector/market?
43. Have you ever co-founded a startup prior to this current project?
44. What are your personal motivations for creating your own company?
45. What is your main set of skills?
46. What problem is your company trying to address?
47. What is the solution you are offering and who is it targeted to?
48. How many customers have you already interviewed in order to understand their needs?
49. How much would you earn each year, if you could sell your solution to ALL the existing customers of the entire market?
50. What is the market share of the current market leader?
51. What is the market potential?
52. What are the alternatives to your solution currently on the market?
53. What is your competitive advantage?
54. Describe your business model and what makes it viable and original
55. What was your amount of turnover over the last three months?
56. What is your current weekly turnover growth rate?
57. What is your conversion rate?

58. What's your pricing and how did you determine it?
59. Who are your current and future partners and why?
60. What are your achievements and actions so far?
61. If you have a prototype, please provide a link to a demonstration
62. Is your solution based on one of the following technologies?
63. What is your product roadmap for the next 3 months?
64. What is your business roadmap for the next 3 months?
65. What are your ambitions for the company on the long term (3+ years), middle term (1 year), short term (3 months)?
66. How much financing do you need to make your project happen?
67. What is the amount of money you already raised?
68. Describe : How much have you already raised and from whom?
69. When will you reach the breakeven point?
70. Have you identified any new regulatory changes that might create opportunities or issues for your activity? If so, please describe those changes.
71. Have you identified any negative environmental/societal externalities created by your company?
72. Will addressing these externalities be one of your priorities?
73. How do you define your positive social AND environmental impact?
74. What are your motivations and what do you want to bring to the Incubator?

APPENDIX F: EVALUATION CRITERIA

Below I report the set of evalaution criteria and associated scoring rubric filled out by human and AI evaluators.

- Among the following 4 criteria, does this team have the necessary skills and connections? Entrepreneurship, Sector, Job, Mentors.
    1. 1 or less
    2. 2 out of 4
    3. 3 +
- Does this project address a concrete user pain-point?
    1. Inexistant, poorly identified and poorly expressed
    2. Blurry, Partly identified and somewhat clearly expressed
    3. Concrete, clearly identified and well expressed
- Does this project bring a feasible solution to the pain-point?
    1. No
    2. Partially
    3. Yes
- Does the team have a clear understanding of competition?
    1. They display a poor analysis of their competitors and substitutes
    2. They display a moderately clear analysis of their competitors and substitutes
    3. They display a very detailed analysis of their competitors and substitutes
- Does this project have a clear competitive advantage based on some key differentiating factor?
    1. No unique key asset, skill or product feature
    2. Somewhat unique, non-imitable key asset, skill or product feature
    3. Unique, non-imitable key asset, skill or product feature
- Does this project address current or expected regulation affecting the business opportunity?
    1. Known regulation or regulatory changes are not addressed
    2. No need to address regulation, or regulation is partially addressed
    3. Known regulation or regulatory changes are addressed
- Does this project have a viable business model?
    1. Poor idea about value creation and monetization

2. Some idea about value creation and monetization
3. Understand value creation and monetization

- Has the project gained traction at this time?
  1. No prospects
  2. Have prospects. A prospective customer, or prospect, is a person or organization interested in making a purchase, with financial resources required, and the power to make purchasing decisions.
  3. Have paying customers
- What is their business development status today? (partners, supply chain, production, operations)
  1. Value architecture has not been imagined at all
  2. Value architecture is being explored, but has not been validated
  3. Value architecture is designed and validated
- Is the projected sales volume going to be excellent?
  1. Team do not demonstrate an understanding of how to create sales, they have a ”nice to have” product
  2. Team shows some understanding of how to sell with limited value proposition for the clients
  3. Team demonstrates a good and REALISTIC understanding of how to create sales
- What is their technical product development status today? (TRL = technological readiness level)
  1. The product is at its research phase : basic principles observed and formulated concept (TRL 1-3)
  2. The product is at the development phase : it is being tested and validated with a few customers (TRL 4-6)
  3. The product is at the deployment phase : Fully functional products available to a large pool of customers (TRL 7-9)
- Would a VC or BA invest in this project?
  1. No
  2. Maybe
  3. Yes
- Does the project have the necessary financing in place for the next 12 months?
  1. Financing is in place for less than 3 months of expenses

2. Financing is in place for 3 to 6 months of expenses
3. Financing is in place for more than 6 months of expenses

- Have the founders demonstrated that they can execute at high speed?
  1. No indication they execute fast
  2. Some indications they execute fast
  3. Strong indications that they execute fast
- Does this team have the sufficient motivation and ambition?
  1. Only lifestyle ambitions, low ambition to scale
  2. They are only here by opportunity but could be doing something else
  3. This is the project of their life

APPENDIX G: PROMPT SPECIFICATIONS

This appendix reports the text of the seven prompt specifications used to generate AI evaluations: the baseline and the six variations tested for robustness in Section 5.1. All seven specifications share the criterion-by-criterion scoring rubric for X1 through X15, reported in appendix F, task restrictions and output instructions, (boxes below) and venture's application formatted as a sequence of question-answer pairs listed in appendix E.
The specifications differ only in the system-level instructions describing the evaluation task, reported in Boxes G.1 through G.6, or, for the Application Question Order specification, in the order in which the application's question-answer pairs are supplied to the model.

Output Instructions

```
OUTPUT INSTRUCTIONS (very important):
For EACH criterion X1..X15, return:
1) "Score": the single best score (1, 2, or 3) that you judge most
   appropriate for this criterion, based on the application and the
   scoring guidance.
2) "P(1)", "P(2)", "P(3)": your confidence distribution over the
   three possible scores, reflecting how well each score fits the
   application for that criterion.
- Each probability must be between 0 and 1.
- The three probabilities must sum to 1.
- Exactly one of the three probabilities must be strictly greater
   than the other two.
3) "justification": the top 3 question IDs from the application (e.g
   ., "Q12") whose answers most support your chosen Score for that
   criterion.
- Return exactly 3 items.
- Each item must match the pattern "Q<number>".

Return ONLY the JSON object (no markdown, no commentary).
```

Task Restrictions

```
You must base your evaluation ONLY on:
1) the venture application provided by the user (questions Q1..Q74
   and answers), and
2) knowledge relevant for the evaluation of a venture that would have
    been available up to the application's submission date shown in
   the application after question "Round.1.Creation.Date".
```

```
Do NOT use any knowledge or events after that date.

The application may be written in French (and may include French
    names/terms). That is expected. You may interpret it in its
    original language. Regardless of language, you must follow the
    output requirements below.

You will evaluate the venture on 15 criteria (X1..X15) using the
    following criteria and scoring guidance:
```

Box E.1. Role-Based (Baseline)

```
You are an evaluator at a startup incubator called HEC Incubator. You
     assess whether a venture should be admitted to the final stage
    of the selection process.

[Task restrictions]

[Criteria and scoring rubric, X1-X15]

[Output Instructions]

Here is the venture application to evaluate:

[Application text]
```

Box G.2. Zero-Shot

```
Please read the text of the following venture application to HEC
    Incubateur and evaluate the venture.

[Task restrictions]

[Criteria and scoring rubric, X1-X15]

[Output Instructions]

Here is the venture application to evaluate:
```

```
[Application text]
```

Box G.3. Few-Shot

```
Please read the text of the following venture application to HEC
    Incubateur and evaluate the venture.

[Task restrictions]

[Criteria and scoring rubric, X1-X15]

To calibrate your scoring, consider the following worked examples.
    Each illustrates how the scoring guidance applies to a different
    criterion, using a different venture. The same evaluative
    standard - grounding each score in concrete, verifiable evidence
    from the application rather than in aspirational claims - applies
     across all 15 criteria, not only the three illustrated here.

For Criterion X12 ("Would a VC or BA invest in this project?"):
"A biotech venture with a patented diagnostic platform, two signed
    pilot contracts with regional hospitals, and a founding team
    combining a PhD in molecular biology with a former medtech
    commercial director" = 3 (Concrete IP protection, paying or
    piloting customers, and a team with both technical and commercial
     credibility together provide the evidence base a VC or BA would
    require to underwrite the risk.)

For Criterion X2 ("Does this project address a concrete user pain-
    point?"):
"A wearable-device application stating that 'many people feel anxious
     in social situations and want tools to help them cope,' without
    specifying which population, which situations, or what evidence
    supports the claim" = 2 (The application points to a broad,
    plausible human experience but does not narrow it to a specific,
    verifiable population or triggering context, so the pain point is
     only partly identified rather than concrete.)

For Criterion X8 ("Has the project gained traction at this time?"):
"An application describing the product as 'ready to launch,' with no
    named users, no waitlist, and no pilot agreement or letter of
    intent from any prospective customer" = 1 (Traction requires
    observable prospects or customers; a readiness claim alone
```

provides no evidence that anyone beyond the founders has engaged with the product.)

Apply this same evidentiary standard - favoring specific, checkable facts over stated ambition - when scoring all 15 criteria.

[Output Instructions]

Here is the venture application to evaluate:

[Application text]

Box G.4. Chain of Thought

Please read the text of the following venture application to HEC Incubateur and evaluate the venture.

[Task restrictions]

[Criteria and scoring rubric, X1-X15]

Before assigning scores, work through the application using the following explicit analytical steps. Use these steps only to structure your internal reasoning; do not include the reasoning itself in your output.

Step 1 - Team and execution capacity: Identify the founders' relevant sector, entrepreneurial, and professional experience, and any named mentors or advisors, to assess whether the team can execute on the stated plan.

Step 2 - Problem and solution fit: Determine whether the application identifies a concrete, specific user pain point and whether the proposed solution plausibly addresses it, distinguishing verified claims from aspirational statements.

Step 3 - Market and competitive positioning: Assess the stated market size, competitive landscape, and differentiating factor, checking whether claims of uniqueness are supported by specific, non-generic evidence.

```
Step 4 - Commercial and financial traction: Evaluate the evidence of
    business model viability, current traction (customers, revenue,
    partnerships), and financial runway relative to the venture's
    stage.

Step 5 - Integration: Combine the conclusions from Steps 1-4 to
    assign a Score and probability distribution for each of the 15
    criteria, ensuring each justification cites the specific question
     IDs that informed the relevant step.

[Output Instructions]

Here is the venture application to evaluate:

[Application text]
```

Box G.5. Self-Consistency

```
Please read the text of the following venture application to HEC
    Incubateur and evaluate the venture.

[Task restrictions]

[Criteria and scoring rubric, X1-X15]

Before producing your final answer, internally perform three
    independent evaluation passes over the application, each with a
    distinct analytical focus. Use the three passes only to inform
    your single final answer; do not report them separately in your
    output.

Pass 1 - Evidentiary pass: Score each criterion based strictly on
    verifiable facts stated in the application (customers, contracts,
     credentials, dates, figures), disregarding unsubstantiated
    claims of ambition or intent.

Pass 2 - Comparative pass: Score each criterion by benchmarking the
    venture against the typical profile of an early-stage applicant
    to a university-affiliated incubator at the same stage of
    development.
```

```
Pass 3 - Skeptical pass: Score each criterion by actively searching
    for gaps, inconsistencies, or overstated claims in the
    application, applying additional scrutiny before conferring the
    higher end of the scale.

For each criterion, set your final Score to the majority outcome
    across the three passes (the score at least two of the three
    passes converge on). If all three passes disagree, use Pass 1 (
    the evidentiary pass) as the tie-breaker. Set "P(1)", "P(2)", "P
    (3)" to reflect how consistently the three passes converged on
    each possible score: strong agreement should produce a sharply
    peaked distribution, and disagreement should produce a flatter
    one.

[Output Instructions]

Here is the venture application to evaluate:

[Application text]
```

Box G.6. Reflection on Search Trees

```
Please read the text of the following venture application to HEC
    Incubateur and evaluate the venture.

[Task restrictions]

[Criteria and scoring rubric, X1-X15]

For each criterion, before assigning a score, internally weigh the
    strongest argument in favor of a high score (3) against the
    strongest argument in favor of a low score (1), using only
    evidence present in the application. Use this pro/con reasoning
    only to inform your single final answer; do not report it in your
     output.

Argument for a high score: Identify the strongest concrete evidence
    in the application supporting a favorable assessment on this
    criterion.
```

```
Argument for a low score: Identify the strongest concrete evidence,
    gap, or omission in the application supporting an unfavorable
    assessment on this criterion.

Strength assessment: Determine which argument is better supported by
    specific, verifiable evidence in the application, rather than by
    the confidence or length of the applicant's stated claims.

Confidence assessment: Let the relative strength of the two arguments
     determine your probability distribution across P(1), P(2), and P
    (3): a clear-cut argument should produce a sharply peaked
    distribution, while closely matched competing arguments should
    produce a flatter, less confident distribution, with the modal
    score still reflecting your best final judgment.

[Output Instructions]

Here is the venture application to evaluate:

[Application text]
```