Title

The Uneven Impact of Generative Artificial Intelligence on Entrepreneurial Performance: Evidence from a Field Experiment in Kenya

Journal

Management Science

ISSN

0025-1909

Authors

Otis, Nicholas G

Clarke, Rowan

Delecourt, Solène

et al.

Publication Date

2026-07-10

DOI

10.1287/mnsc.2024.06909

The Uneven Impact of Generative AI on Entrepreneurial Performance: Evidence from a Field Experiment in Kenya

Nicholas G. Otis
Berkeley Haas

Rowan Clarke
Harvard Business School

Solène Delecourt
Berkeley Haas

David Holtz
Columbia Business School

Rembrand Koning
Harvard Business School

Scalable and low-cost AI assistance has the potential to improve firm decision-making and economic performance, particularly in emerging markets. However, running a business involves a wide range of open-ended problems, making it unclear whether and how recent advances in AI can help business owners around the world make better decisions. In a field experiment with Kenyan entrepreneurs, we evaluated the impact of AI advice on small business revenues and profits by randomizing access to a GPT-4-powered AI business assistant. While we are unable to reject the null hypothesis of no *average* treatment effect on firm revenues and profits, we find that the effect for entrepreneurs who were low-performing at baseline is over 0.20 standard deviations lower than for initial high performers. Subsample analyses show that low performers did nearly 10% worse due to the AI assistant, whereas high performers may have benefited by over 15%. This differential impact does not appear to result from differences in the questions posed to the AI or the advice it provided, but rather from the advice entrepreneurs chose to implement. More broadly, these results show that generative AI is already capable of impacting real-world business performance—though in uneven and sometimes unexpected ways.

1 Introduction

Since the launch of ChatGPT in November 2022, there has been an explosion of research on generative AI and its potential economic implications (The White House, 2022; Agrawal, Gans, and Goldfarb, 2023; Eloundou et al., 2024; Goldfarb, 2024). Much of this recent work is driven by the belief that conversations with large language models (LLMs) can help people learn and develop new skills (Choi et al., 2023; Mollick and Mollick, 2022), and in work contexts, can improve firm performance and growth (Brynjolfsson, Li, and Raymond, 2025; Dell'Acqua et al., 2023; Noy and Zhang, 2023; Peng et al., 2023; Kumar et al., 2023). Given the substantial variation in worker and firm productivity, both within and between countries, the emergence of almost zero marginal-cost generative AI "assistants" has the potential to radically improve the productivity and performance of everyone, from the thousands of CEOs running publicly listed companies in the United States to the hundreds of millions of entrepreneurs running small and medium-sized businesses in developing economies (McAfee, Rock, and Brynjolfsson, 2023; Björkegren, 2023).

Consistent with the optimism currently surrounding generative AI (AI hereafter), recent experiments show that conversing with AI and receiving AI assistance causes workers to write better business text, including press releases, ad copy, consulting memos, and customer support messages, more quickly (Brynjolfsson, Li, and Raymond, 2025; Dell'Acqua et al., 2023; Doshi and Hauser, 2023; Noy and Zhang, 2023; Chen and Chan, 2023). However, it remains unclear how the benefits of such AI feedback generalize to the broader set of tasks that firms engage in. Beyond these text-based tasks, firms must also manage employees, raise capital, pilot new initiatives, run advertising strategies, price their services, react to competitors, and decide which of these and myriad other tasks to focus their efforts on (Chandler, 1977). The sheer multitude of tasks involved in running a business greatly increases the complexity of effectively learning how to improve business performance, especially for small businesses and entrepreneurs in emerging markets, which often lack the team, training, and advisors that drive learning (Kim, 2024; Hanna, Mullainathan, and Schwartzstein, 2014). And even when they have access to human advisors and mentors, entrepreneurs often struggle to decide which tasks they should ask for help with, how to formulate effective questions to get useful

feedback, and how to interpret and act upon the advice they receive (Bryan, Tilcsik, and Zhu, 2017; Camuffo et al., 2020; Agrawal, Gans, and Stern, 2021; Dimitriadis and Koning, 2022).

It is possible that these challenges are exacerbated when seeking advice from AI systems. Entrepreneurs may be reluctant to ask questions of an AI assistant, rather than another person (Lebovitz, Lifshitz-Assaf, and Levina, 2022). Without good questions, even advanced AI systems may fail to provide relevant or useful guidance, and even when an entrepreneur's question is well-crafted, it is unclear how useful AI advice will be in practice. Given that many business tasks are tacit and may not yet be codified as text—especially tasks in emerging markets—AI tools may lack the relevant training data to provide helpful answers (Autor, 2014; Tao et al., 2023). AI systems can also produce ineffective or overconfident "solutions" that, if implemented, could worsen rather than improve firm performance (Dell'Acqua et al., 2023; Ji et al., 2023). Moreover, entrepreneurs may lack the judgment, complementary knowledge, skills, or resources needed to select and implement helpful AI-generated suggestions while avoiding harmful ones (Brynjolfsson and Hitt, 2000; Brynjolfsson, Rock, and Syverson, 2021; Agrawal, Gans, and Stern, 2021). However, if entrepreneurs are able to formulate good questions, if AI provides useful suggestions, and if entrepreneurs can effectively use their judgment to act on this guidance, AI advice has the potential to help entrepreneurs, especially in emerging markets.

In this paper, we explore the potential of AI to impact firm performance in a randomized controlled trial (RCT) with 640 Kenyan small business entrepreneurs, half of whom were randomly assigned to receive access to a GPT-4-powered AI assistant over WhatsApp. Our experiment lasted approximately five months, from late May 2023 to early November 2023. We are unable to reject the null hypothesis that, on average, access to generative AI had no effect on business performance relative to the control. Although our overall treatment effect estimate is statistically indistinguishable from zero, this result masks substantial treatment effect heterogeneity with respect to pre-treatment firm performance. We estimate that the effect of treatment for entrepreneurs with below-median baseline performance (henceforth "low performers") is over 0.20 standard deviations lower than for entrepreneurs with above-median baseline performance ("high performers"). Subsample analysis shows that receiving

the AI assistant led to an approximately 0.08 decrease in performance for low performers, whereas high performers may have experienced performance gains exceeding 0.15 relative to the control.

To better understand the mechanisms driving this heterogeneity, we analyze the questions entrepreneurs asked, the AI's responses, and post-treatment survey descriptions of business changes. We find no evidence that performance differences stem from variation in the number or types of questions asked, the advice generated, or the overall likelihood of acting on AI suggestions. Instead, high performers worked with the AI to identify and implement tailored, context-specific changes, while low performers disproportionately adopted generic strategies, such as lowering prices and increasing advertising, which may have reduced revenues and raised costs.

Our results make three primary contributions to the literature on the economic impact of AI. First, by examining entrepreneurial decision making, we provide evidence on whether generative AI can deliver benefits in complex, interconnected contexts beyond the well-defined tasks emphasized in prior work (Sato et al., 2023). Extending results from studies of narrowly defined work tasks (Noy and Zhang, 2023; Dell'Acqua et al., 2023), our findings suggest that while AI can generate a wide range of potentially useful business advice, the benefits depend critically on which advice users choose to implement (Dell'Acqua et al., 2023; Wiles and Horton, 2024). Second, we contribute evidence from an emerging market setting that has been largely absent from recent empirical studies of generative AI (Björkegren, 2023). Third, we show that although generative AI offers a scalable, low-cost way to extend business mentorship and training globally (McKenzie, 2021), realizing these benefits broadly will require improvements in both AI systems and how entrepreneurs and workers are trained to use them.

2 Literature Review and Motivation

Recent experimental studies show that generative AI can automate and enhance a wide range of well-defined knowledge work tasks. Noy and Zhang (2023) find that US white-collar workers who received generative AI training completed writing tasks roughly 40%

faster while improving output quality by 18%. Similarly, Dell'Acqua et al. (2023) show that BCG consultants using GPT-4 developed strategic plans 25% faster and received 40% higher quality ratings, with particularly large gains among initially weaker performers. The benefits extend to creative work: Doshi and Hauser (2023) show that AI-assisted participants wrote more creative short stories, and Meincke et al. (2024) find that AI-generated product ideas outperformed those produced by students at an elite U.S. university.[1] Together, these results point to substantial economic potential for generative AI in structured knowledge tasks.

However, these task-level findings raise questions about how AI performs in "messier" real-world organizational settings. As Noy and Zhang (2023) note, real-world tasks often involve vague objectives and require workers to exercise initiative in deciding where to apply AI. Although some workplaces resemble structured experimental environments—such as the standardized call center studied by Brynjolfsson, Li, and Raymond (2025)—most organizations operate with much less structure. In these settings, workers must decide whether and where to deploy AI assistance, and this discretion matters: Dell'Acqua et al. (2023) show that AI can have strongly positive or negative effects depending on the tasks to which it is applied. As a result, variation in the way workers and firms deploy AI may shape both its average impact and its distributional consequences.

This emphasis on discretion echoes a long literature on decision support systems, which shows that differences in judgment, incentives, and organizational context often limit the benefits of machine-generated assistance. Early optimism in the 1970s and 1980s that computer-based interactive dialogue would improve decision-making (Keen and Scott Morton, 1978; Sprague Jr, 1980) gave way to mixed empirical results, with limited gains in many organizational settings (Sharda, Barr, and McDonnell, 1988). These outcomes frequently reflected how users incorporated—or failed to incorporate—tool outputs into their decision processes. Consistent with this pattern, recent studies show that reliance on AI predictions can sometimes hinder performance (Kim et al., 2024; Agarwal et al., 2023), particularly when users lack the judgment to selectively implement useful recommendations.

To understand AI's broader impact, we therefore need evidence from settings that combine task complexity with user discretion. Entrepreneurship offers a particularly relevant

[1]Both studies also note that AI assistance led to more homogeneous outputs.

context. As Lazear (2004) argues, entrepreneurs perform an unusually wide range of tasks, making them well suited for assessing how AI's effects generalize beyond controlled experiments. Entrepreneurs must decide whether to use AI, for which tasks, and how to act on its output, requiring judgment and contextual knowledge (Agrawal, Gans, and Goldfarb, 2022). Indeed, Noy and Zhang (2023) suggest that when AI is applied to open-ended, context-dependent tasks, it may amplify differences in talent rather than substitute for expertise. These dynamics make entrepreneurship—especially small-scale business activity—a natural setting for studying AI's distributional effects (Khanna, 2014).

Testing AI in emerging markets extends this literature. Limited representation of these contexts in training data may lead LLMs to provide less useful or inaccurate advice (Björkegren, 2023; Tao et al., 2023). More broadly, differences between developing and advanced economies—in firm tasks, worker human capital, and decision-making constraints—may produce distinct effects of AI on performance. The skill frontier dynamics documented by Doshi and Hauser (2023) illustrate this point: high performers benefited little from AI because they already operated near the frontier. In developing economies, where many workers are farther from this frontier, patterns of AI adoption and impact may differ substantially. Given that over 80% of the world's population lives outside developed economies, understanding AI's performance in these settings is central to assessing its global economic potential. The policy implications are potentially significant: if AI can substitute for scarce business expertise in resource-constrained environments, it may provide a scalable tool for upskilling entrepreneurs and workers where returns are highest (McKenzie, 2021).

3 Experimental Design and Analysis

3.1 Study Setting and Sample Recruitment

Kenya is a well-suited setting to test a WhatsApp-based generative AI intervention because it combines a large informal business sector (Kenya National Bureau of Statistics, 2023) with widespread use of digital communication technologies—over 90% of Kenyan adults own a mobile phone (Kamau et al., 2024), and most use WhatsApp for text communication

(Shengale, 2025). This level of digital access makes a WhatsApp-based intervention feasible and allowed us to run an experiment across Kenya (Appendix Figure A1 shows the geographic distribution of respondents), while highlighting a potential boundary condition for deploying similar tools in other emerging market contexts.

The experiment proceeded in three stages: recruitment and pre-treatment surveys, random assignment to either a generative AI business assistant or control group, and four post-treatment survey waves (see Appendix Figure A2). First, we recruited small and medium-size business (SMB) entrepreneurs through an online advertisement and collected three waves of pre-treatment performance data (see Appendix Subsection C.1). The final sample included 640 Kenyan SMB entrepreneurs from 44 of the country's 47 counties, with the largest shares from Nairobi (133), Kiambu (53), and Nakuru (47). The average entrepreneur was nearly 26 years old; most had at least some college education, and only 10% reported no college. Women comprised 32% of the sample, consistent with lower female business ownership rates in Kenya (The World Bank Group, 2018).

The sample spans a wide range of business types, including fast food vendors, poultry farmers, and cybershop owners. The most common sectors were Food and Beverage (27%), Clothing (21%), and Agriculture (19%). The median entrepreneur had run their oldest business for one year or less and had modest pre-treatment performance: average monthly profit was 19,550 Kenyan Shillings, approximately $137.[2]

3.2 Experimental Interventions

Following the third pre-treatment survey, participants were randomly assigned to either the treatment group, which received access to the generative AI business assistant, or the control group, which received access to publicly available business training guides described below. To improve statistical precision, randomization was stratified by quartiles of pre-treatment business performance and by gender. The two groups were balanced across baseline characteristics and performance measures (Table A1).[3]

[2]We convert to USD using the exchange rate near the midpoint of our experiment, July 31st, 2023, using Wise's KES to USD historical rate calculator.

[3]Post-treatment attrition was similar across conditions. We classify participants as having attritted if they did not complete the final survey, even if they completed earlier post-treatment surveys. Of the 640

Before gaining access to their assigned intervention, participants in the AI condition completed a short (5–10 minute) online introduction to the AI assistant. This introduction stated that the chatbot was an AI system (not a human), that its information might be outdated, and that its advice could sometimes be inaccurate. To maintain comparability between conditions, control participants received a similar set of introductory materials focused on the control group guides rather than the AI tool (see Appendix C).

3.2.1 Treatment (The AI Business Assistant)

The AI-powered business assistant was built on GPT-4, a large language model released by OpenAI in March 2023 (OpenAI, 2023). To ensure accessibility, the assistant was deployed on WhatsApp, which is widely used across Kenya. Access to the assistant was restricted so that it responded only to phone numbers belonging to participants in the treatment group.

Interacting with the assistant differed from directly using ChatGPT in three main ways. First, we implemented a custom *system prompt*—a set of initial instructions provided to the model—to tailor its responses to the context of Kenyan SMBs. For example, when asked about raising funds, ChatGPT might suggest raising venture capital, whereas our assistant instead recommended contextually relevant options such as borrowing from family and friends, joining informal investment clubs ("chama"), or using community-based lending institutions. The complete system prompt is provided in Appendix Subsection C.2.

Second, the assistant was instructed to provide three to five numbered recommendations in response to each question, each with a brief explanation. For example, an entrepreneur asking how to respond to increased competition might receive suggestions such as (1) expanding their menu, (2) improving customer service, and (3) introducing a loyalty program.

Third, the assistant supported follow-up interactions. Entrepreneurs could request more details on any recommendation by replying with its corresponding number. In the example above, entering "3" ("loyalty program") generated more specific suggestions, such as using punch cards or low-cost digital tools to track points. This structure encouraged entrepreneurs to explore options before deciding whether to implement them. An example interaction with

participants, 11 attritted from the AI group and 7 from the control group; the 1.2 percentage point difference in response rates is not statistically significant ($p = 0.35$).

the AI assistant is shown in Figure 1.

Entrepreneurs used the assistant to discuss a wide range of topics, including motivating employees, allocating finite resources during business expansion, poultry farming practices, and managing bankruptcy. In Appendix D, we present two detailed conversation logs: a restaurant owner facing increased competition and a dairy salesperson responding to supplier price increases. In both cases, entrepreneurs described their challenges, evaluated multiple options, and returned for additional advice as new issues emerged.

3.2.2 Control (Business Training Guides)

Two considerations guided the design of the control group. First, the AI assistant can provide several types of benefits, including personalized advice and structured reflection, making it difficult to define a precise counterfactual. We therefore sought a control condition that encouraged participants to reflect on their business without using an AI tool.

Second, as in many studies in emerging markets, business performance is measured through self-reports, raising concerns about demand effects. If one group receives a business-focused intervention while the other did not receive anything, differences in reported outcomes may partly reflect reporting expectations rather than actual performance changes. We also wanted to send reminders to promote engagement; providing reminders to only one group would confound reminder effects with the effect of the assistant.

To address these concerns, we designed a control condition that closely paralleled the treatment group. Control participants received links to a series of business training guides over WhatsApp. These guides, developed by the International Labour Organization for SMB entrepreneurs in emerging markets, focus on business development and growth strategies (ILO, 2014). Participants received links to the first three PDFs from a nine-part series, along with a link to the remaining materials (see Subsection C.3). Because prior evidence shows that delivering static training content via text message has limited effects on firm performance (Mehmood, 2023), we treat this condition as a placebo control — an interpretation we support by documenting that participants have minimal engagement with the guides.

3.3 Outcomes

To measure firm performance, we follow standard approaches in emerging market research by collecting multiple self-reported performance measures over multiple survey waves (McKenzie, 2012) and constructing indices to reduce measurement noise (Kling, Liebman, and Katz, 2007). In each of the three pre-treatment and four post-treatment waves, we collected self-reported weekly and monthly profits and revenues (De Mel, McKenzie, and Woodruff, 2009). For example, when eliciting weekly [monthly] profits, respondents were asked: *Now we would like to ask you about the* ***total profits*** *across all of your businesses in the* ***last 7 [30] days***. *What was the total income the business earned during the last 7 [30] days after paying all expenses including the wages of employees, but not including any income you paid yourself.* Parallel questions were asked for weekly and monthly revenue (see Appendix E for a full description).[4] In total, we collected 4,434 observations for each of the four components of the business performance index (2,226 from the AI-assistant group and 2,208 from the control group; summary statistics are reported in Table A2).

Our primary outcome is a business performance index aggregating weekly and monthly profits and revenues. This composite index reduces noise common in self-reported data and is widely used in survey-based research (Kling, Liebman, and Katz, 2007) and emerging market studies (Campos et al., 2017; Anderson and McKenzie, 2022). Aggregating outcomes also mitigates concerns about multiple hypothesis testing. We construct the standardized index following Kling, Liebman, and Katz (2007): each measure is winsorized (at the 99% or 95% level, depending on the specification), standardized using the control group's post-treatment mean and standard deviation, and then averaged across index components within each survey wave. Differences in the performance index between treatment and control participants provide an aggregate measure of the AI assistant's impact on firm performance in standard deviation units, such that a estimated treatment effect of 0.1 (-0.1) would indicate a 0.1 (-0.1) standard deviation increase (decrease) in firm performance.

[4]These measures, along with additional prespecified outcomes examining experimental mechanisms, were registered following the third pre-treatment wave at `https://osf.io/64kws/files/osfstorage/64e386ae62e5ee196a3ab9c9`.

3.4 Estimation Strategy

We estimate the average treatment effect of the generative AI assistant on the business performance index using an OLS regression:

$$y_{i,t} = \beta_0 + \beta_1 \mathbb{1}(\text{AI}_i) + \delta y_{i,-t} + \mathbf{X}_i + \tau_t + \theta_i + \varepsilon_{i,t} \tag{1}$$

where $y_{i,t}$ is the business performance index for participant i in post-treatment period t. $\mathbb{1}(\text{AI}_i) \in \{0, 1\}$ is an indicator equal to 1 if participant i was assigned to the AI assistant and 0 otherwise. The coefficient β_1 captures the average treatment effect of the AI assistant relative to the control group. This regression also controls for pre-treatment values of the outcome variable ($y_{i,-t}$), which can increase statistical power in randomized experiments compared to a difference-in-difference approach (McKenzie, 2012). In addition, the specification includes time-period fixed effects τ_t, strata fixed effects θ_i, and a broader set of controls $\mathbf{X}_i$ selected using the double-LASSO procedure of Belloni, Chernozhukov, and Hansen (2014). Candidate covariates for selection using the LASSO procedure included age, business sector, number of children, education, frequent ChatGPT use, Kenyan county, oldest business age, personality measures (agreeableness, conscientiousness, extroversion, openness, and neuroticism), and pre-treatment management practices.

To estimate heterogeneous treatment effects for our standardized business performance index, we use the following equation:

$$y_{i,t} = \beta_0 + \beta_1 \mathbb{1}(\text{AI}_i) + \beta_2 \text{Het}_i + \beta_3 (\mathbb{1}(\text{AI}_i) \times \text{Het}_i) + \delta y_{i,-t} + \mathbf{X}_i + \tau_t + \theta_i + \varepsilon_{i,t} \tag{2}$$

where Het_i represents a prespecified dimension of heterogeneity, and β_3 captures the differential effect of AI assistant access along that dimension. We examine heterogeneous treatment effects with respect to each of the following three pre-registered variables: (*i*) initial business performance, (*ii*) gender, and (*iii*) prior ChatGPT use. We operationalize initial business performance as a median split of the standardized performance index across the three pre-treatment periods, and frequent ChatGPT use is defined as reporting use of ChatGPT at least once per week prior to the experiment.

4 Results

4.1 Impacts of the AI Assistant on Firm Performance

Our first set of results examines the impact of access to the generative AI assistant on the standardized business performance index. We find no statistically significant average treatment effect of access to the AI assistant on firm performance (winsorized at the 99% level) relative to control, as shown in Estimate A in Figure 2 and Table 1 ($\beta = 0.05$ standard deviations (sd), $p = 0.32$, 95% CI = $[-0.05, 0.14]$). In other words, in the context of small and medium-sized Kenyan businesses, we do not find any evidence that providing access to a generative AI assistant improves firm performance *on average*.

4.2 Treatment Effect Heterogeneity

Building on recent studies showing that AI assistance especially helps lower performing workers (Dell'Acqua et al., 2023; Noy and Zhang, 2023; Peng et al., 2023; Brynjolfsson, Li, and Raymond, 2025), we next estimate heterogeneous treatment effects by splitting our sample into two groups, based on the pre-treatment values of the same performance index we use as our primary outcome variable.[5] We refer to entrepreneurs who had below-median pre-treatment performance as *low performers* and those who had above median pre-treatment performance as *high performers*.

Estimates B and C in Figure 2 and Table 1 show the estimated treatment effect within the low- and high-performer sub-samples. For low performers, access to the AI assistant reduced average business performance relative to the control group ($\beta = -0.08$ s.d., $p < 0.01$, 95% CI = [-0.13, -0.03]). Because our index is standardized against the standard deviation of the post-treatment control group (Kling, Liebman, and Katz, 2007), we can contextualize the magnitude of our effect by looking at the mean and standard deviation of revenues and profits for this group. For example, Table A2 shows that average post-treatment monthly revenue in the control group is 51,392 Kenyan Shillings (USD $361) with a standard deviation of

[5]We also prespecified testing for heterogeneity with respect to gender and prior use of ChatGPT. We report the results of these tests in Figure A3. We find no evidence of treatment effect heterogeneity with respect to either of these dimensions.

59,353 Shillings. Multiplying this standard deviation by our effective size estimate for the monthly revenue index component yields an estimated revenue decline of 3,957 shillings per month ($28) – a 8% drop in revenue. We observe similar estimates for the other index components, which yield negative treatment effect estimates among initially low-performing entrepreneurs of around 10%.

For high performers, we find suggestive evidence that AI assistance improved performance relative to control ($\beta = 0.16$ s.d., $p = 0.07$, 95% CI = $[-0.01, 0.34]$). We can multiply the same standard deviation of monthly revenue by our estimated impact on the monthly revenue index component estimate to get an estimated increase in revenue of 9,493 Kenyan Shillings ($67), equal to a relative increase of about 18%. Again, we observe suggestively positive estimates across the other performance index components.

Finally, building on these subgroup analyses, Estimate D in Figure 2 and Table 1 directly compares the impact of the AI assistant on initially low and high performers using the heterogeneous treatment effects model in Equation 2. We find strong evidence that the effect of AI is different for low and high performers. Initially, low-performing entrepreneurs experienced a 0.23 sd lower causal effect ($p = 0.01$, 95% CI = $[-0.41, -0.06]$) of the AI assistant compared to high performers.[6]

4.3 Robustness

We assess robustness along three dimensions. First, we examine whether heterogeneous treatment effects by baseline performance are driven by outliers or reflect broader distributional shifts. Figure A4 plots residualized performance distributions for initially low- (Panel A) and high-performing firms (Panel B). In Panel A, the control distribution lies to the right of the treated, indicating higher performance among control participants, while in Panel B this pattern reverses. These shifts in the empirical cumulative distribution functions suggest that the heterogeneous effects reflect broad changes in profits and revenues rather than a small number of extreme observations.

[6]Given this uneven effect, and especially the negative performance effect we find for low performers, in Appendix F we discuss ethical considerations and the importance of policy equipose in designing and interpreting experiments.

Second, we assess sensitivity to alternative specifications. Figure A5 shows that the core pattern—a null average treatment effect masking heterogeneity by baseline performance—holds when winsorizing outcomes at the 95% rather than 99% level and when restricting the sample to participants who completed all post-treatment surveys. Across all of our performance index specifications, our most conservative estimate of treatment effect heterogeneity is 0.18 s.d. ($p < 0.05$, 95% CI $[0.02, 0.35]$). Figures A6 and A7 show qualitatively similar patterns for individual components of the performance index, and A8 presents robustness across different empirical specifications. Appendix G shows that results are unlikely to be driven by spillovers between participants, and Appendix Figure A9 plots survey-wave–specific estimates, showing that divergence in performance emerges only post-treatment.

Finally, we address concerns about multiple hypothesis testing. We apply the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) to control the false discovery rate across three families of outcomes: (*i*) main performance effects, (*ii*) heterogeneous treatment effects, and (*iii*) survey-based mechanism measures. Adjusted p-values are reported in Table A3, and our key results have similar levels of statistical significance after correction.

5 Unpacking Mechanisms

To understand why low and high performers exhibit different performance effects, we adopt an abductive approach and use rich text data from AI interactions and post-treatment surveys to identify potential mechanisms (Pillai, Goldfarb, and Kirsch, 2024). As discussed in the literature review, most prior studies of generative AI focus on structured tasks. In contrast, participants in our setting had substantial discretion over how and when to use AI assistance: they could apply AI output to many tasks and had to exercise judgment about whether to implement its advice. To assess how this discretion shaped AI's impact, we structure the mechanism analysis in three parts. First, we examine whether low- and high-performing entrepreneurs differ in the number or types of questions they ask. Second, we test whether the AI's responses differ across groups. Third, we analyze whether entrepreneurs act on the AI's advice and whether low and high performers select different

types of suggestions to implement.

5.1 Data and Analytical Methods

We use three data sources to explore the mechanisms behind the heterogeneous performance effects.[7] First, we analyze the set of messages sent to the AI assistant. Entrepreneurs sent 4,810 messages, with 88% of treated entrepreneurs sending at least one message, and these entrepreneurs sent an average of 17 questions. Second, we use the corresponding responses from the AI assistant. This system typically returned three to five recommendations per response, as shown in Figure 1, and the average response from the AI assistant was approximately 167 words long.

Third, to measure how entrepreneurs used AI advice, we analyze open-text responses from post-treatment survey questions asking entrepreneurs to describe changes to their (i) products, (ii) services, (iii) business processes, (iv) the most impactful business changes they made in the past 30 days, and (v) any other new business activities. We combine these five responses into a single text object per entrepreneur (see Subsection H.1 for details). Unlike the AI interaction data, business-change text is available for both treated and control participants. As shown in Table 2, among treated participants, initial performance status does not predict availability of question or answer text, and among all participants, neither treatment status nor performance predicts availability of business-change text, mitigating selection concerns in text-based comparisons.

The mechanisms we test—whether entrepreneurs ask different questions, implement different changes, and whether AI advice causally shapes those changes—are conceptually clear but empirically challenging for two reasons. First, text varies along many semantic dimensions, so detecting differences requires high-dimensional semantic representations rather than single-variable content coding. Second, AI advice is endogenous: the advice depends on entrepreneurs' questions, so naive correlations between AI content and business changes risk confounding. We address these challenges with recent advances in text analysis, machine

[7] We also prespecified alternative dependent variables based on post-treatment multiple-choice survey questions to test a host of potential mechanisms. This analysis, described in Appendix Subsection E.2 and Figure A10, examines differences in areas such as management practices and stigma in advice seeking. Across the seven mechanisms tested, we do not find any evidence of statistically significant effects.

learning, and causal inference. We represent text with embeddings, use supervised machine learning to test for systematic semantic differences, and combine embeddings with Pearl's backdoor criterion to measures the impact of AI advice on business changes.

Our approach follows recent management research using embeddings to represent complex semantic relationships while remaining computationally tractable (Cao, Koning, and Nanda, 2023; Goldberg and Srivastava, 2022). We use text embeddings to map the text data described above into high-dimensional vectors that quantitatively capture semantic relationships: semantically similar phrases (e.g., "reduce prices" and "offer discounts") have similar vectors even if they share no words (Mikolov et al., 2013). We generate 3,072-dimensional embeddings using OpenAI's `text-embedding-3-large` model for each of our three types of text data: the entrepreneurs' questions to the AI assistant, the AI's responses, and entrepreneurs' post-treatment open-text descriptions of business changes. For each entrepreneur, we concatenate all texts of a given type into a single text object before generating the embedding.

To test whether semantic content differs systematically across groups, we frame the analysis as a supervised prediction problem, asking whether embedding vectors can predict an entrepreneur's pre-treatment performance or treatment assignment. For example, if low performers discuss discounting-related ideas more frequently, the corresponding embedding components should enable classification above chance; conversely, failure to predict is consistent with there being no meaningful semantic differences. We implement this using random forest classifiers, which handle high-dimensional data and capture nonlinear interactions (Breiman, 2001). For each model, we perform 100 Monte Carlo iterations in which the data is split 60/40 into training and test sets. Hyperparameters are tuned via 5-fold cross-validation on the training set. We report the mean accuracy and 95% empirical intervals, defined as the 2.5th and 97.5th percentiles of the accuracy distribution across iterations; see Appendix Subsection H.2 for details.

Finally, to help estimate the causal effect of AI advice on business changes, we apply Pearl's backdoor criterion (Pearl, 2009). Entrepreneur characteristics could influence both questions and subsequent business changes, creating spurious correlations between AI advice and reported changes. For example, an entrepreneur already planning a marketing campaign

may ask about advertising, receive advertising-related advice, and then report an ad campaign. Because the AI has no information about entrepreneurs beyond their messages, conditioning on question content blocks these backdoor paths (see Figure A11). Operationally, we represent questions, AI advice, and business-change text with embeddings. Across all dimension embeddings d we then estimate:

$$\text{BusinessChange}_{i,d} = \beta\, \text{AIAdvice}_{i,d} + \gamma\, \text{Question}_{i,d} + \varepsilon_{i,d}, \tag{3}$$

where controlling for $\text{Question}_{i,d}$ closes the confounding path from entrepreneur characteristics to both AI advice and business changes.

5.2 Testing for Differences in Question Asking

We first test whether heterogeneous treatment effects could stem from differences in how low- and high-performing entrepreneurs interact with the AI assistant. We find no evidence of such differences. Panel A of Table A5 shows that a similar number of low and high performers sent at least one message (144 vs. 137), that entrepreneurs ask a comparable number of questions (18.7 vs. 15.5 on average), submit a similar number of numeric follow-ups (5.35 vs. 4.56), and write questions of similar length (13.4 vs. 14.5 words).[8]

To test whether question *content* differs, we use the 3,072-dimensional vector representing each entrepreneur's full set of questions to train a random forest to predict pre-treatment performance status. If low-and high-performers asked systematically different questions, these embeddings should be predictive. Instead, the model achieves a mean accuracy of 52.8% across 100 Monte Carlo train/test splits (Table 3; 95% empirical interval [0.45,0.61]), similar to the 51.4% no-information benchmark. To further examine this result, we analyze the topics entrepreneurs asked about. Since many messages are numeric follow-ups, fragments, or non-business content (e.g., pressing 3" or replying y"), we restrict all topic analysis to the 1,450 substantive business messages identified by an independent pair of human coders (see Appendix H.1 for details). Appendix Figure A14 shows that low and high performers

[8]Median message counts are also nearly identical (9 vs. 8 messages for low and high performers, respectively).

asked about similar topics. Overall, we find no evidence of differences in either the volume or content of questions asked.

5.3 Testing for Differences in AI Advice

Although entrepreneurs ask similar questions, the AI might still respond differently across groups. We test this possibility by examining both the amount and content of advice received. Panel B of Table A5 shows that because entrepreneurs ask a similar number of questions, they also receive a similar volume of advice. Moreover, the length of responses from the AI is nearly identical, with an average length of 168 words for initially low-performing entrepreneurs and 166 words for high performers.

Next, we test whether advice content differs by training a random forest on the embeddings that represent each entrepreneur's full set of AI-generated responses to predict pre-treatment performance. The model achieves a mean accuracy of 45.3% (Table 3; 95% empirical interval [0.37,0.53]), which falls below the no-information benchmark. Appendix Figure A15 similarly shows that AI responses to low and high performers cover comparable topics. Thus, we find no evidence the AI provides systematically different advice to the two groups.

5.4 Do Entrepreneurs Follow AI-Generated Suggestions?

We next test whether entrepreneurs act on AI-generated suggestions. Using the backdoor empirical strategy described in Subsection 5.1, we regress embedding dimensions of reported business changes on corresponding dimensions of AI advice, clustering standard errors by entrepreneur and embedding dimension (Table 4). Model 1 shows a strong correlation: a one standard deviation increase in an AI advice embedding dimension is associated with a 0.26 standard deviation increase in the corresponding business-change dimension ($p < 0.001$).

To address confounding, Model 2 controls for question embeddings; the estimate remains large and significant at 0.124 standard deviations ($p < 0.001$, 95% CI $[0.104, 0.144]$), implying that roughly half the raw correlation may reflect the causal effect of AI advice. Model

3 confirms robustness to entrepreneur fixed effects.[9] We find no evidence that low- and high-performers differ in their propensity to follow AI advice. Model 4 interacts advice and question embeddings with a high-performer indicator; neither interaction is statistically significant. Models 5 and 6 estimate the relationship by baseline performance group and yield nearly identical coefficients. Thus, both groups follow AI advice to a similar extent.

5.5 Do High and Low Performers Implement Different Changes?

Because low and high performers ask similar questions, receive similar advice, and both act on that advice, a remaining explanation for heterogeneous treatment effects is that they differ in how they select and implement AI-generated suggestions. We test this by analyzing self-reported text describing entrepreneurs' business changes.

Using embedding-based random forests trained on business-change text from all treated and control entrepreneurs with available data, we find consistent evidence of differential implementation. Model 3 in Table 3 predicts whether an entrepreneur is a low or high performer with a mean accuracy of 60.4% (95% empirical interval $[0.54, 0.66]$), exceeding the 50.9% no information benchmark. Model 4 predicts treatment status with a mean accuracy of 54.7% (95% empirical interval [0.50,0.60]), suggestively above the 50.5% benchmark. Most relevant to the interpretation of our heterogeneous effects, Model 5 jointly predicts performance level and treatment status (four categories) with a mean accuracy of 33.3% (95% empirical interval $[0.27, 0.40]$) versus a 25.4% baseline.[10] These results indicate that AI access and baseline performance jointly shape the types of changes entrepreneurs implement.

To interpret these differences, we develop a novel empirical strategy to identify business changes that best reflect the average semantic differences between treated and control entrepreneurs. We compute group-specific "treatment effect embeddings" by first averaging the business-change embeddings separately for treated and control entrepreneurs within each performance group, and then taking the element-wise difference between these two averages—analogous to a simple difference in means, but in a high-dimensional semantic space. We then identify treated–control pairs of business-change text whose embedding differences

[9] A complementary bag-of-words style analysis yielding similar conclusions is reported in Table A6.
[10] The lower baseline reflects prediction across four categories rather than two.

are most similar to these treatment effect embeddings using cosine similarity. By selecting examples from the top 1% most similar pairs, we highlight the types of changes that best characterize the average treatment effect for each group (see Appendix H).

The resulting examples reveal a clear contrast. Panel A of Table 5 shows that treated high performers frequently describe working explicitly with the AI assistant to implement targeted, business-specific changes (e.g., managing electricity outages or introducing new revenue streams), with substantial variation across cases—consistent with the absence of heterogeneous effects in our prespecified, more generic mechanism measures (Figure A10). Panel B shows that treated low performers rarely mention the AI and more often report generic strategies such as discounting or advertising, which are less differentiated and may be less likely to generate sustained performance gains (Carlson, 2023).

This pattern motivates our preferred interpretation: high-performing entrepreneurs are more likely to select and implement AI advice that is tailored to their specific business context, whereas low-performing entrepreneurs tend to adopt more generic recommendations. While alternative explanations—such as differences in implementation effort, tolerance for complexity, or time horizons—may also contribute, the specificity and business relevance of high performers' changes, together with explicit references to AI guidance, point to differential selection and implementation of contextually appropriate advice as a plausible mechanism underlying our heterogeneous treatment effects.

We further test this interpretation using quantitative text-based measures. We construct indicators for (i) AI-related engagement (explicit references to the AI assistant), (ii) specificity (use of uncommon words appearing in five or fewer entrepreneurs' texts), and (iii) generic strategies (mentions of discounting or advertising). Figure 3 shows that treated entrepreneurs are 6.2 percentage points more likely to reference AI use ($p < 0.001$, 95% CI $[3.32, 9.11]$), an effect driven by high performers (10.3 points, $p < 0.001$) rather than low performers (1.8 points, $p = 0.19$), with a significant between-group difference ($p < 0.01$). Treated high performers use about 1.2 times as many uncommon words than control ($p < 0.05$, 95% CI $[1.03, 1.41]$), while low performers show no such increase. Conversely, treated low performers are significantly more likely to report discounts or advertising than control, with a difference of 11.9 points, $p < 0.05$, 95% CI $[1.67, 22.08]$, with no

corresponding effect among high performers. Additional details are provided in Appendix Subsection H.6.

5.6 The Role of Judgment in Open-Ended Settings

Beyond explaining the uneven performance effects in our study, our mechanism analysis suggests a more general framework for understanding when AI might produce unequal impacts. Prior generative AI experiments have largely focused on well-defined tasks, where the sign of AI's impact on a given task is relatively consistent across users and benefits are often largest for those who previously struggled (Dell'Acqua et al., 2023; Noy and Zhang, 2023; Brynjolfsson, Li, and Raymond, 2025). In these settings, the task structure reduces the need for users to exercise judgment about which AI suggestions to follow. In contrast, entrepreneurs in our setting face many open-ended tasks simultaneously and receive a mixture of helpful and unhelpful suggestions. Our text analysis shows that both low and high performers receive a mixture of generic and tailored AI advice, but high performers are better at selecting and implementing the more promising, context-specific advice, causing a reversal of the pattern observed in prior work. It's possible that this dynamic is especially pronounced in emerging markets, where limited representation in AI training data may reduce the average quality of AI-generated advice (Björkegren, 2023; Tao et al., 2023). Figure A16 provides a stylized illustration of how variation in advice quality and user judgment can generate heterogeneous treatment effects such as those we observe.

6 Discussion

To our knowledge, this study is the first randomized controlled trial exploring the impact of AI on entrepreneurs and firms in a developing economy. Many entrepreneurs in low- and middle-income countries face persistent barriers to accessing personalized feedback, mentorship, and training (Chatterji et al., 2019; Dimitriadis and Koning, 2024; Björkegren, 2023). Our findings show that AI tools can both help and harm entrepreneurs, depending on baseline performance. For high-performing firms, we find suggestive evidence that AI access generated performance gains comparable in magnitude to effective business training pro-

grams, but at a fraction of the cost.[11] At the same time, we find evidence that AI assistance significantly reduced performance among low-performing firms, leading to nearly 10% worse outcomes. These results suggest that while AI has the potential to benefit firms in developing countries (Björkegren, 2023; Lou, Sun, and Sun, 2023), untargeted deployment risks widening existing performance gaps.

Prior studies of AI assistance have largely focused on well-defined tasks with relatively narrow distributions of advice quality (Brynjolfsson, Li, and Raymond, 2025; Dell'Acqua et al., 2023). In contrast, our intervention allowed entrepreneurs broad discretion in how they used AI across many open-ended business tasks, generating a wider distribution of advice quality and implementation outcomes. Our mechanism analysis shows that this setting places greater weight on user judgment: although low and high performers receive similar advice, high performers are more likely to select and implement contextually appropriate, tailored suggestions, while low performers disproportionately adopt generic strategies, such as price cuts or advertising, that can reduce performance. These findings extend prior work by demonstrating that when AI is applied to open-ended decision environments, its impact depends critically on users' selection and implementation capabilities.

6.1 Limitations

Our study has several limitations. First, it examines a single context—small business entrepreneurs in Kenya—over a limited time horizon of five months, and tests a specific AI implementation: a GPT-4–based assistant delivered with a short system prompt and brief user training. Even the highest-performing firms in our sample operate with fewer resources than firms in developed economies, where entrepreneurs often have greater access to education, professional services, and expert advisors. Larger firms may also face fewer capital constraints, enabling more transformative AI-enabled investments. While this limits direct generalizability, hundreds of millions of firms worldwide resemble the microenterprises in our study, underscoring the relevance of our setting and the need for broader evidence across firm size, sector, and geography.

[11] Similar to text-message interventions (Fabregas et al., 2024), our AI assistant cost a few dollars per participant, compared to hundreds of dollars for most training programs (McKenzie, 2021).

Second, although our study duration is longer than most prior AI experiments (cf. Brynjolfsson, Li, and Raymond, 2025), longer-run impacts may differ. If low-performing firms implement changes that take time to yield returns, our estimates may understate benefits; conversely, if generic strategies such as discounting or advertising prove increasingly harmful over time, we may understate negative effects. Our experiment also took place in mid-2023, when familiarity with AI tools was still limited in many developing contexts. As exposure and learning increase, both usage patterns and effects may evolve.

Third, our AI assistant relied on a relatively short system prompt and minimal training for users. Further prompt iteration, domain-specific fine-tuning, or additional user training could plausibly amplify benefits for high performers and reduce harms for low performers. Heterogeneous impacts may also reflect biases in GPT-4's training data, suggesting that models trained on more representative data could perform better in emerging-market contexts (Tao et al., 2023).

Finally, while our abductive mechanism analysis is most consistent with differences in judgment and ability driving heterogeneous effects, we cannot fully rule out alternative explanations. As in many emerging-market studies (De Mel, McKenzie, and Woodruff, 2009), outcomes are self-reported and may reflect reporting behavior rather than real changes, despite our use of an active control group. In addition, although our sample size is comparable to similar experiments (McKenzie, 2021), statistical power remains limited (McKenzie, 2025), and some patterns—particularly among high performers—could reflect sampling variability.

6.2 Implications

Our findings point to several implications for research and policy. First, understanding why entrepreneurs select and implement different types of advice is critical. Low-performing entrepreneurs appear highly responsive to AI advice, even when it harms performance, raising questions about whether better guidance or training could improve outcomes. The strong correlation and causal link between AI advice content and subsequent business changes (Table 4) highlights both the promise and risk of AI systems: when advice is generic or poorly contextualized, users may be led astray. Even among high performers, it remains unclear whether AI's impact is near the frontier of attainable returns (Doshi and Hauser,

2023) or could be enhanced through complementary resources such as access to capital or implementation support.

In developing economies, our results caution against untargeted AI deployment. Without careful design or complementary training, AI tools may disproportionately benefit already-successful firms while harming those that are struggling, potentially widening economic disparities. At the same time, the fact that most entrepreneurs follow at least some AI advice suggests enormous potential scale. As with business training programs (McKenzie, 2021), future research should focus on how AI systems can be better designed, targeted, and bundled with other interventions to promote inclusive growth (George, McGahan, and Prabhu, 2012). Domain-specific AI tools—for example, focused on agriculture, finance, or market research—may reduce the need for users to filter large volumes of mixed-quality advice.

Finally, our results have implications beyond developing economies. The central mechanism we identify—that AI's impact depends on users' judgment in selecting and implementing advice—likely applies broadly wherever AI supports open-ended decision making rather than narrowly defined tasks. As AI capabilities expand and are applied to more domains, judgment and contextual understanding may become increasingly important rather than less. Understanding how firms and workers adapt their skills, decision processes, and interactions with AI (Lebovitz, Lifshitz-Assaf, and Levina, 2022; Boussioux et al., 2023; Girotra et al., 2023; Hui, Reshef, and Zhou, 2023)—including how they communicate context and adjust to evolving models (Jahani et al., 2024)—will be essential for anticipating equilibrium outcomes as both AI and users learn.

By studying AI's impact on the open-ended, everyday decisions that define real business operations (Harrison and List, 2004), our experiment shows how AI's effects can diverge sharply from those observed in controlled, task-specific settings. Much like the distinction between *in vitro* and *in vivo* research in medicine, evidence from controlled settings is crucial for isolating mechanisms of action, while evidence from more open-ended, less controlled environments is essential for understanding how those mechanisms operate in practice. These findings underscore the importance of evaluating AI across diverse contexts and populations—not only to assess its economic potential, but to ensure that its deployment expands opportunity rather than deepening existing gaps.

References

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining human expertise with artificial intelligence: Experimental evidence from radiology." Tech. rep., National Bureau of Economic Research.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2022. *Prediction machines, updated and expanded: The simple economics of artificial intelligence*. Harvard Business Press.

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb. 2023. "Do we want less automation?" *Science* 381 (6654):155–158.

Agrawal, Ajay, Joshua S Gans, and Scott Stern. 2021. "Enabling entrepreneurial choice." *Management Science* 67 (9):5510–5524.

Anderson, Stephen J and David McKenzie. 2022. "Improving business practices and the boundary of the entrepreneur: A randomized experiment comparing training, consulting, insourcing, and outsourcing." *Journal of Political Economy* 130 (1):157–209.

Autor, David. 2014. "Polanyi's paradox and the shape of employment growth." NBER Working Paper 20485.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-dimensional methods and inference on structural and treatment effects." *Journal of Economic Perspectives* 28 (2):29–50.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57 (1):289–300.

Björkegren, Daniel. 2023. "Artificial Intelligence for the Poor: How to Harness the Power of AI in the Developing World." Foreign Affairs.

Boussioux, Leonard, Jacqueline N Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R Lakhani. 2023. "The Crowdless Future? How Generative AI Is Shaping the Future of Human Crowdsourcing." *The Crowdless Future* .

Breiman, Leo. 2001. "Random forests." *Machine learning* 45 (1):5–32.

Bryan, Kevin A, András Tilcsik, and Brooklynn Zhu. 2017. "Which entrepreneurs are coachable and why?" *American Economic Review* 107 (5):312–316.

Brynjolfsson, Erik and Lorin M Hitt. 2000. "Beyond computation: Information technology, organizational transformation and business performance." *Journal of Economic Perspectives* 14 (4):23–48.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. "Generative AI at work." *The Quarterly Journal of Economics* :qjae044.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2021. "The productivity J-curve: How intangibles complement general purpose technologies." *American Economic Journal: Macroeconomics* 13 (1):333–372.

Campos, Francisco, Michael Frese, Markus Goldstein, Leonardo Iacovone, Hillary C Johnson, David McKenzie, and Mona Mensmann. 2017. "Teaching personal initiative beats traditional training in boosting small business in West Africa." *Science* 357 (6357):1287–1290.

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

Cao, Sam Ruiqing, Rembrand Koning, and Ramana Nanda. 2023. "Sampling Bias in Entrepreneurial Experiments." Management Science.

Carlson, Natalie A. 2023. "Differentiation in microenterprises." *Strategic Management Journal* 44 (5):1141–1167.

Chandler, Alfred D. 1977. *The visible hand*. Harvard university press.

Chatterji, Aaron, Solène Delecourt, Sharique Hasan, and Rembrand Koning. 2019. "When does advice impact startup performance?" *Strategic Management Journal* 40 (3):331–356.

Chen, Zenan and Jason Chan. 2023. "Large language model in creative work: The role of collaboration modality and user expertise." *Available at SSRN 4575598* .

Choi, Jun Ho, Oliver Garrod, Paul Atherton, Andrew Joyce-Gibbons, Miriam Mason-Sesay, and Daniel Björkegren. 2023. "Are LLMs Useful in the Poorest Schools? theTeacherAI in Sierra Leone." *arXiv preprint arXiv:2310.02982* .

De Mel, Suresh, David J McKenzie, and Christopher Woodruff. 2009. "Measuring microenterprise profits: Must we ask how the sausage is made?" *Journal of development Economics* 88 (1):19–31.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." SSRN 4573321.

Dimitriadis, Stefan and Rembrand Koning. 2022. "Social skills improve business performance: Evidence from a randomized control trial with entrepreneurs in Togo." *Management Science* 68 (12):8635–8657.

———. 2024. "Networking Frictions: Evidence from entrepreneurial networking events in Lomé." Management Science (forthcoming).

Doshi, Anil R and Oliver Hauser. 2023. "Generative artificial intelligence enhances creativity." Available at SSRN 4535536.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs are GPTs: Labor market impact potential of LLMs." Science.

Fabregas, Raissa, Michael Kremer, Matthew Lowes, Robert On, and Giulia Zane. 2024. "Digital information provision and behavior change: Lessons from six experiments in East Africa." Tech. rep., National Bureau of Economic Research.

George, Gerard, Anita M McGahan, and Jaideep Prabhu. 2012. "Innovation for inclusive growth: Towards a theoretical framework and a research agenda." *Journal of management studies* 49 (4):661–683.

Girotra, Karan, Lennart Meincke, Christian Terwiesch, and Karl T Ulrich. 2023. "Ideas are dimes a dozen: Large language models for idea generation in innovation." Available at SSRN 4526071.

Goldberg, Amir and Sameer B Srivastava. 2022. "How Can AI Enrich Our Understanding of Organizational Culture."

Goldfarb, Avi. 2024. "Pause artificial intelligence research? Understanding AI policy challenges." *Canadian Journal of Economics/Revue canadienne d'économique* .

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2014. "Learning through noticing: Theory and evidence from a field experiment." *The Quarterly Journal of Economics* 129 (3):1311–1353.

Harrison, Glenn W and John A List. 2004. "Field experiments." *Journal of Economic Literature* 42 (4):1009–1055.

Hui, Xiang, Oren Reshef, and Luofeng Zhou. 2023. "The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market." SSRN 4527336.

ILO. 2014. *Start and Improve Your Business (SIYB)*. Geneva. URL `https://www.ilo.org/empent/areas/start-and-improve-your-business/WCMS_192062/lang--en/index.htm`. Accessed: 2025-05-18.

Jahani, Eaman, Benjamin S Manning, Joe Zhang, Hong-Yi TuYe, Mohammed Alsobay, Christos Nicolaides, Siddharth Suri, and David Holtz. 2024. "As generative models improve, we must adapt our prompts." *University of California, Berkeley* .

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. "Survey of hallucination in natural language generation." *ACM Computing Surveys* 55 (12):1–38.

Kamau, Paul, Laura Barasa, Gedion Onyango, and Oscar Otele. 2024. "Summary of results: Afrobarometer Round 10 survey in Kenya, 2024." Survey summary, Institute for Development Studies, University of Nairobi & Afrobarometer. URL `https://www.afrobarometer.org/wp-content/uploads/2024/10/Summary-of-results-Kenya-Afrobarometer-R10-template-Eng-16oct24-for-upload.pdf`.

Keen, Peter GW and Michael S Scott Morton. 1978. "Decision support systems: an organizational perspective." *(No Title)* .

Kenya National Bureau of Statistics. 2023. "Economic Survey 2023: Popular Version." URL `https://www.knbs.or.ke/wp-content/uploads/2023/09/2023-Economic-Survey-Popular-Version.pdf`.

Khanna, Tarun. 2014. "Entrepreneurship in emerging markets: Contextual intelligence for the study of two thirds of the world's population." In *Multidisciplinary insights from new AIB fellows*, vol. 16. Emerald Group Publishing Limited, 221–238.

Kim, Hyunjin. 2024. "The value of competitor information: Evidence from a field experiment." Management Science (forthcoming).

Kim, Hyunjin, Edward L Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2024. "Decision authority and the returns to algorithms." *Strategic Management Journal* 45 (4):619–648.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz. 2007. "Experimental analysis of neighborhood effects." *Econometrica* 75 (1):83–119.

Kumar, Harsh, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. "Math Education with Large Language Models: Peril or Promise?" Available at SSRN 4641653.

Lazear, Edward P. 2004. "Balanced skills and entrepreneurship." *American Economic Review* 94 (2):208–211.

Lebovitz, Sarah, Hila Lifshitz-Assaf, and Natalia Levina. 2022. "To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis." *Organization science* 33 (1):126–148.

Lou, Bowen, Hongshen Sun, and Tianshu Sun. 2023. "GPTs and labor markets in the developing economy: Evidence from China." *Available at SSRN 4426461* .

McAfee, A, D Rock, and E Brynjolfsson. 2023. "How to Capitalize on Generative AI." *Harvard Business Review* 101 (6):42–48.

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics* 99 (2):210–221.

———. 2021. "Small business training to improve management practices in developing countries: re-assessing the evidence for 'training doesn't work'." *Oxford Review of Economic Policy* 37 (2):276–301.

———. 2025. "Designing and analysing powerful experiments: practical tips for applied researchers." *Fiscal Studies* .

Mehmood, Muhammad Zia. 2023. "Short Messages Fall Short for Micro-Entrepreneurs: Experimental Evidence from Kenya." .

Meincke, Lennart, Karan Girotra, Gideon Nave, Christian Terwiesch, and Karl T Ulrich. 2024. "Using large language models for idea generation in innovation." *The Wharton School Research Paper Forthcoming* 9:2024.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* .

Mollick, Ethan R and Lilach Mollick. 2022. "New modes of learning enabled by ai chatbots: Three methods and assignments." *Available at SSRN 4300783* .

Noy, Shaked and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science* 381:187–192.

OpenAI. 2023. "GPT-4 Technical Report."

Pearl, Judea. 2009. *Causality*. Cambridge university press.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. "The Impact of AI on Developer Productivity: Evidence from Github Copilot." *arXiv preprint:2302.06590* .

Pillai, Sandeep Devanatha, Brent Goldfarb, and David Kirsch. 2024. "Lovely and likely: Using historical methods to improve inference to the best explanation in strategy." *Strategic Management Journal* 45 (8):1539–1566.

Sato, Megan Kinniment Lucas Jun Koba, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. 2023. "Evaluating Language-Model Agents on Realistic Autonomous Tasks." arXiv: 2312.11671.

Sharda, Ramesh, Steve H Barr, and James C McDonnell. 1988. "Decision support system effectiveness: a review and an empirical test." *Management science* 34 (2):139–159.

Shengale, Ram. 2025. "WhatsApp Statistics 2025: Usage Trends, Demographics & More." URL `https://wanotifier.com/whatsapp-statistics/`.

Sprague Jr, Ralph H. 1980. "A framework for the development of decision support systems." *MIS quarterly* :1–26.

Tao, Yan, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2023. "Auditing and Mitigating Cultural Bias in LLMs." arXiv: 2311.14096.

The White House. 2022. ""The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America." Technical Report.

The World Bank Group. 2018. "Enterprise Surveys: What Businesses Experience - Kenya 2018 Country Profile." URL `https://www.enterprisesurveys.org/content/dam/enterprisesurveys/documents/country-profiles/Kenya-2018.pdf`.

Wiles, Emma and John J Horton. 2024. "More, but Worse: The Impact of AI Writing Assistance on the Supply and Quality of Job Posts." .

Figures

Figure 1: Example Interaction with the AI Assistant

I have fast food joint that I run just adjacent to a matatu terminus. There are several such food joints at that particular place and it's fair to say competition is stiff. How can I overcome the competitive business environment and be ahead of the rest of the pack?

Thank you for sharing about your business. Operating a fast food joint in a highly competitive environment like a matatu terminus can indeed be challenging, but there are several strategies you could consider to stand out:

1. **Differentiate Your Menu**: Offer unique or special dishes that aren't available at the other joints. This could be traditional Kenyan foods with a twist, or it could involve using locally sourced ingredients that others don't use. The benefit is that customers will associate your place with these special offerings and choose you over competitors.
2. **Improve Customer Service**: Prioritize excellent customer service to make people feel welcome and appreciated when they visit your joint. This might involve training staff on how to interact with customers positively, quickly serving orders, and handling any complaints promptly and professionally.
3. **Loyalty Program**: Implement a loyalty program where customers earn points for every purchase which they can later redeem for free meals or discounts. It encourages repeat business as customers will want to accumulate more points.
4. **Community Involvement**: Engage in local community events or sponsor local initiatives such as sports teams or school activities. This raises your profile within the community and makes people more likely to support your business.
5. **Cleanliness & Hygiene**: Ensure high standards of cleanliness both inside the kitchen and in the dining area because hygiene matters greatly in food businesses.

Enter a number to learn more about one of these points on "improving competitiveness" or ask a new question.

3

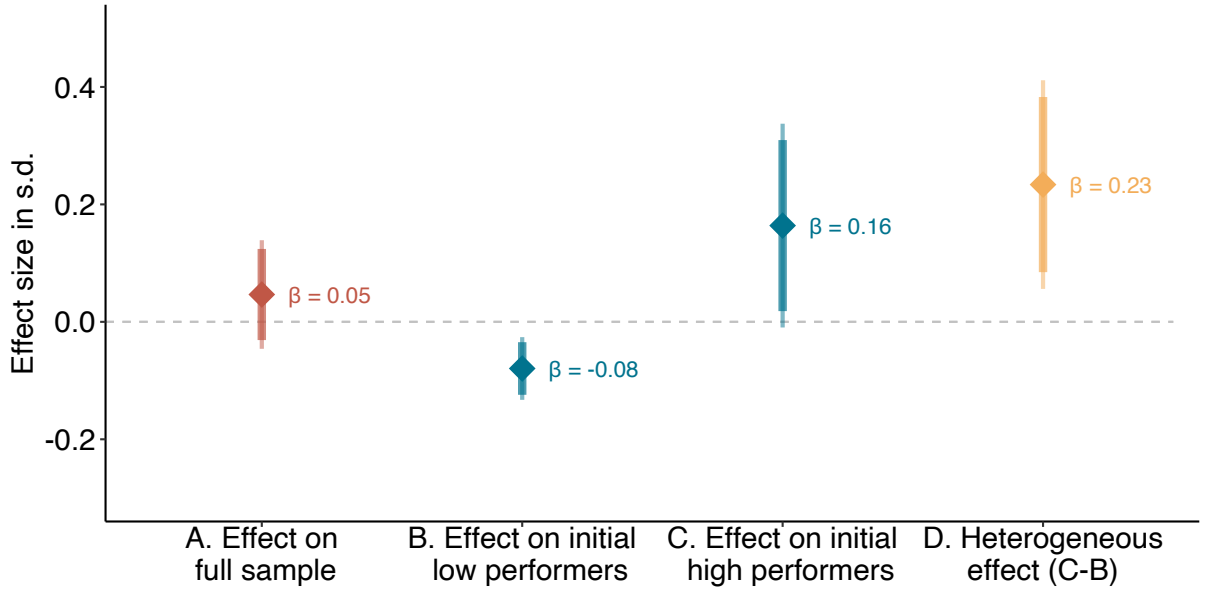
Great, let's delve into how you can implement a **Loyalty Program** in your fast food joint:

1. **Point-Based System**: Implement a simple point-based system where customers earn points for every shilling spent. Once they accumulate a certain number of points, they can redeem them for free meals or discounts. You could use simple tools like punch cards or digital apps that are inexpensive to manage.
2. **Special Offers for Regular Customers**: Identify your regular customers and offer them exclusive deals or discounts. This not only rewards their loyalty but also encourages them to keep coming back.
3. **Birthday/Anniversary Rewards**: Offer special treats or discounts to customers on their birthdays or anniversaries if they share this information with you. This personalized touch can make customers feel valued and appreciated.
4. **Referral Program**: Encourage your loyal customers to bring in new clients by giving them incentives such as a discount or free meal when someone they refer makes their first purchase.
5. **Feedback Reward**: Encourage customer feedback by offering small rewards for completing surveys about the quality of food and service at your joint.

Remember, the goal of any loyalty program is to make your customers feel valued so that they continue patronizing your business over others. Enter a number to learn more about one of these points on "loyalty programs" or ask a new question.

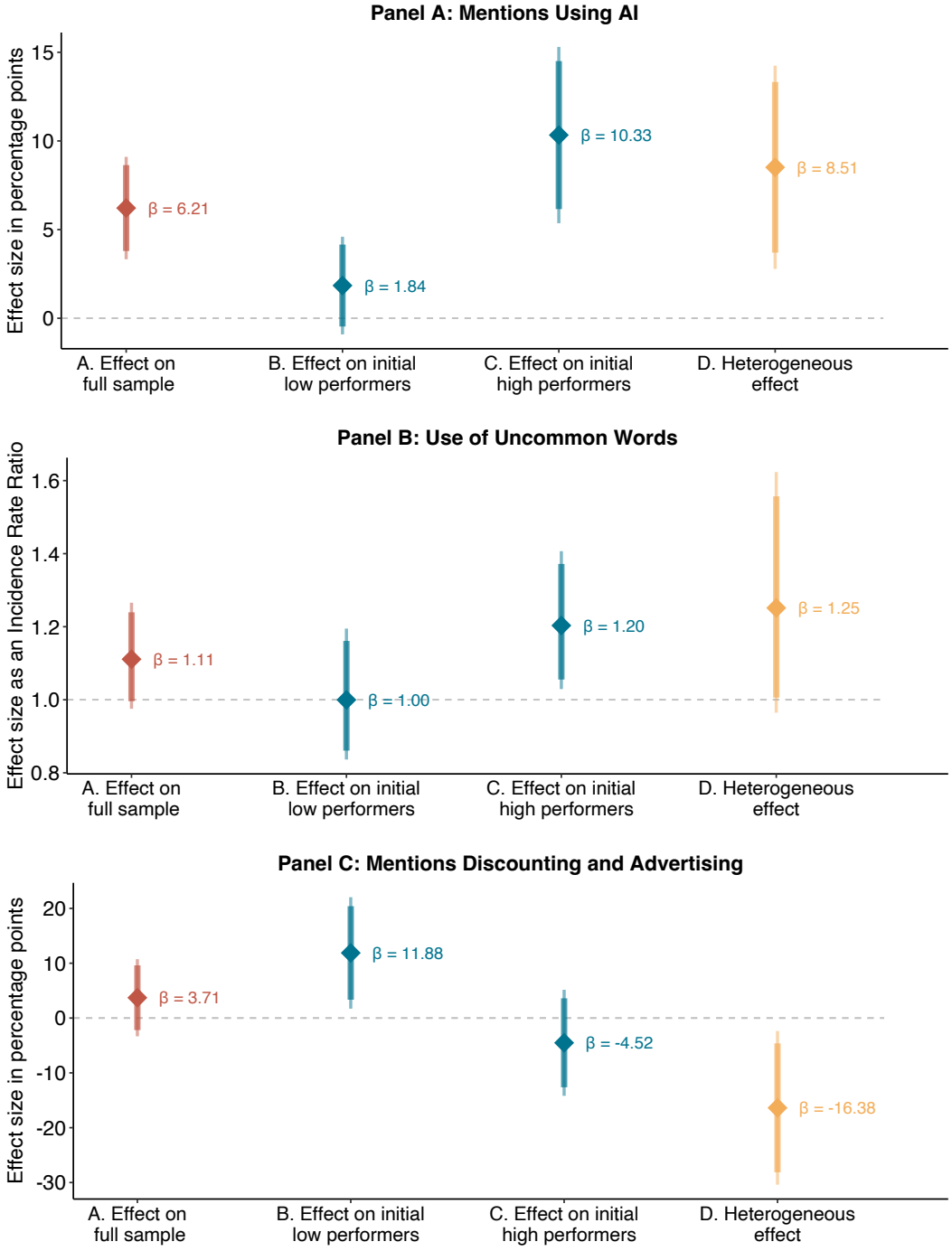
This figure presents a rendering of an entrepreneur-AI interaction in WhatsApp. The text is original (including the typos). A "matatu" refers to a minibus in Kenya.

Figure 2: Effects of the AI Assistant on Business Performance



The impact of access to the AI assistant on the standardized business performance index. Effects are estimated controlling for pre-treatment business performance, time fixed effects, strata fixed effects, and other covariates selected using double-LASSO. Estimate A presents the average treatment effect of the AI assistant. Estimates B and C display results restricted to initially low- and high-performing firms. Estimate D presents the heterogeneous treatment effect. Error bars represent 90% and 95% confidence intervals, with standard errors clustered at the individual level.

Figure 3: The Heterogeneous Impact of the AI Treatment on the Entrepreneurs' Business Changes



These panels quantitatively test the qualitative results in Table 5. Panel A shows that high performers are more likely to mention the AI assistant. Panel B shows that high performers use more uncommon words that may be tailored to their specific business needs. Panel C shows that low performers are more likely to implement generic and potentially costly strategies like lowering prices or investing in ads. Error bars present 90 and 95% confidence intervals. Regressions are at the entrepreneur level. The estimates in Panel A are from the linear probability models in Table A7, the estimates in Panel B are the exponentiated coefficients from the Poisson regressions in Table A8, and the estimates in Panel C are from the linear probability models in Table A7. Error bars represent 90% and 95% confidence intervals, with robust standard errors.

Tables

Table 1: Treatment Effects of AI Assistant Access: Average Effects and Heterogeneity by Pre-Treatment Performance

	Full sample (1)	Below med. (2)	Above med. (3)	Full sample (4)
AI assistant	0.05	-0.08**	0.16†	-0.07*
	(0.05)	(0.03)	(0.09)	(0.03)
AI assistant*(⩾ med.)				0.23*
				(0.09)
Observations	2,514	1,251	1,263	2,514

The impact of access to the AI assistant on firm performance. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Columns A and D report effects across the full sample of entrepreneurs. Columns B and C report effects restricting to participants with above- or below-median pre-treatment performance. Observations refers to the number of performance index observations over the four post-treatment survey waves from the 634 entrepreneurs completing at least one post-treatment performance survey. Effects are estimated controlling for average pre-treatment performance, time fixed effects, strata fixed effects, and covariates selected using double-LASSO. The outcome is winsorized at the 99% level. Standard errors are clustered at the individual level.

Table 2: Text Availability and Word Counts by Treatment Status and Performance Level

	Changes Text		Questions Text		Answers Text	
	Has Text?	Num. Words	Has Text?	Num. Words	Has Text?	Num. Words
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	Poisson	OLS	Poisson	OLS	Poisson
Constant	0.82***	3.83***	0.90***	4.69***	0.90***	7.81***
	(0.03)	(0.08)	(0.02)	(0.12)	(0.02)	(0.13)
AI assistant	0.03	-0.03				
	(0.04)	(0.11)				
$\geqslant$ med.	0.04	0.12	-0.04	0.05	-0.04	-0.12
	(0.04)	(0.11)	(0.04)	(0.17)	(0.04)	(0.18)
AI assistant$\times(\geqslant$ med.)	-0.04	0.11				
	(0.06)	(0.15)				
Subsample	All	All	AI	AI	AI	AI
Observations	634	634	319	319	319	319

Regressions of whether entrepreneurs have available text data and the length of their responses for business changes, questions asked, or AI responses received. This table indicates that neither text availability nor response length is systematically related to treatment assignment or performance level. Word counts are winsorized at the 99% level. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Data are reported for the set of entrepreneurs completing at least one post-treatment survey. Robust standard errors are presented in parentheses.

Table 3: Predictive Accuracy of Text Embeddings

Model	(1)	(2)	(3)	(4)	(5)
Data	*Questions*	*AI Advice*	*Changes*	*Changes*	*Changes*
Prediction	High Performer	High Performer	High Performer	AI Treatment	AI $\times$ High Performer
Classes	2	2	2	2	4
Entrepreneurs	281	281	537	537	537
No Information Rate	0.514	0.514	0.509	0.505	0.254
Mean Accuracy	0.528	0.453	0.604	0.547	0.333
95% Empirical Interval	[0.45, 0.61]	[0.37, 0.53]	[0.54, 0.66]	[0.50, 0.60]	[0.27, 0.40]

Results from random forests fit over high-dimensional text embeddings. In columns 1 and 2, random forests are fit on embeddings from the questions asked and the AI-generated advice, respectively. In columns 3–5, random forests are fit on embeddings from the self-reported business changes. Results are based on 100 Monte Carlo train/test splits. No Information Rate refers to the accuracy achieved by always predicting the majority class, which is the baseline a classifier must exceed to demonstrate predictive signal. 95% empirical intervals are the 2.5th and 97.5th percentiles of the accuracy distribution across iterations.

Table 4: Effects of AI Advice Content on Business Changes

	Changes Embedding Value					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.00	0.00				
	(0.01)	(0.01)				
Answer Embedding Value	0.26***	0.12***	0.12***	0.12***	0.12***	0.13***
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Question Embedding Value		0.35***	0.35***	0.33***	0.33***	0.36***
		(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Answer Embedding Value × (⩾ med.)				0.01		
				(0.01)		
Question Embedding Value × (⩾ med.)				0.03		
				(0.02)		
Performers Subsample	All	All	All	All	Low	High
Entrepreneur Fixed Effects			Yes	Yes	Yes	Yes
Observations	746,496	746,496	746,496	746,496	377,856	368,640

Regressions at the "embedding dimension"×entrepreneur level examine the relationship between AI advice content and business changes using standardized embedding dimensions as observations. Model 1 shows the raw association, while Model 2 applies the backdoor criterion by controlling for question content. Models 3-6 test robustness and show similar causal effects for both high and low performers, indicating that while entrepreneurs implement AI advice, the propensity to do so does not differ by performance level. Figure A11 illustrates how controlling for question content may improve identification by blocking confounding pathways. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors are clustered at entrepreneur and embedding level.

Table 5: Example Business Changes with High Cosine Similarity to Treatment Effect Embedding

Control Examples	AI Examples
A. High Performers	
[. . .] I've started baking honey filled muffins. Honey baked muffins. Home delivery of my products. [. . .]	Use of new technology of AI mentor. By doing more benchmarking on the competitors. Cooking oil, cereals. [. . .]
Adding more products into the business. [. . .] Advertising what am selling. [. . .] Delivering door to door services.	Improved on customer care skills after getting some tips and advice from AImentor and introduction of discounts to customers who purchased the products in bulk. [. . .]
[. . .] I have started selling hair bonnets. [. . .] Delivery of products to customers has really made a difference in my business. Making myself available at any time for my customers. [. . .]	[. . .] I have inserted a small salon beside my cyber [. . .] I got advice from AI mentor on other ways of getting power when there is electricity power black out. [. . .]
[. . .] Cooking quality food. [. . .] Delivering my products to customers. Advertising my business through online. [. . .] Giving my customers carrier bags. [. . .] Selling coffee	Improving the quality of chicks in my chicken roster has brought high quality chicken to my folder... I ma grateful for the Chat AI. [. . .] I have bought new chicks that were recommended by whe WhatsApp AI. [. . .] Reducing slightly the prices of my commodities [. . .]
[. . .] i introduced other kitchen ware sets like cups and plates. Delivery of goods to customers. Engaging clients. Use of media in advertising. Male clothes. [. . .]	[. . .] With the help of AI, I have been able to do more market research and discovered new ways of advertising my products. I have added cellphones, speakers and electric wires [. . .] I have started electronics repairs [. . .] I have started free delivery for Items worth 5000 and above.
B. Low Performers	
Increase in profit in the business. Sales of eggs. Marketing my goods online. Selling of clothes. Transport service. [. . .]	Giving so offers by reducing the selling price. [. . .] Advertising on social media platforms. [. . .]
[. . .] Getting to know what customers wants, precise record keeping and online marketing. [. . .] Reduced owing to reduce debts. [. . .] Providing loan at an interest.	Advertisement and discounts. Offering after sales service. Training of staff. [. . .]
Budgeting and data keeping. [. . .] Increased profits because I can now manage cashflow appropriately without possible losses. Maize. Mpesa services. Fruit selling. [. . .]	Posting online has brought more customers. [. . .] Discount for items above 1000. Offers for goods above 1500. [. . .] Selling shoes in my mtumba warehouse. [. . .] Cereal outside my shop
Consulting my customers. grocery. [. . .] using of technology. selling of cosmetics products. [. . .]	Isuing discounts on some products. Painting my shop as advertising strategy. Discount. Packaging.
[. . .] I have started selling materials for making vitenge. [. . .] Bodaboda service. [. . .] Including mitumbas in shoes selling. Selling credit. Started layering poultry	The free delivery for my products is really working out for me. [. . .] New Imperial leather for men. Hair dressing. [. . .] Rearing and selling chicken

Panel A displays results for initially high-performing entrepreneurs while Panel B displays results for initially low-performing entrepreneurs. Table A9 and Table A10 display extended text results. Highlights correspond to the qualitative differences observed between text from control and treated entrepreneurs.