

WORKING PAPER IDB-WP-01887

Scaling Up Counseling with AI: Evidence from a Nationwide Experiment

Nicolás Ajzenman

Ana Teresa del Toro Mijares

Gregory Elacqua

Catalina Hermosilla

Santiago Veleda

Inter-American Development Bank
Education Division

October, 2026



Scaling Up Counseling with AI: Evidence from a Nationwide Experiment

Nicolás Ajzenman

Ana Teresa del Toro Mijares

Gregory Elacqua

Catalina Hermosilla

Santiago Veleda

Inter-American Development Bank
Education Division

October, 2026



JEL Codes: D91, I23, I25

Keywords: Teachers, teacher shortages, scale-up, behavioral interventions, AI chatbots, generative AI

To explore more IDB publications, visit

<https://publications.iadb.org/en>

Copyright © 2026 Inter-American Development Bank (IDB). This work is subject to a Creative Commons Attribution 4.0 International Public License CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/legalcode>). The terms and conditions indicated in the URL link must be met and the respective recognition must be granted to the IDB.

Any and all disputes arising under this license that cannot be settled amicably shall be resolved in accordance with the following procedure. Pursuant to a notice of mediation communicated by reasonable means by either you or the licensor to the other, the dispute shall be submitted to non-binding mediation conducted in accordance with the World Intellectual Property Organization (WIPO) Mediation Rules. Any dispute that cannot be settled amicably shall be submitted to arbitration pursuant to the United Nations Commission on International Trade Law (UNCITRAL) rules. The use of the IDB name for any purpose other than the respective recognition and the use of the IDB logo are not authorized by this license and require an additional license agreement.

Note that the URL link includes terms and conditions that are an integral part of this license.

The opinions expressed in the work are those of its authors and do not necessarily reflect the views of the IDB, its Board of Executive Directors, or the countries they represent.



Scaling Up Behavioral Interventions with AI: A Nationwide Counseling Experiment*

Nicolás Ajzenman[†] Ana Teresa del Toro Mijares[‡] Gregory Elacqua[§]
Catalina Hermosilla[¶] Santiago Velea^{||}

This version: October 4, 2026

Abstract

A central obstacle to scaling many interventions is implementer quality: as a program expands and the pool of implementers broadens, average quality falls. Conversation-based interventions (counseling, mentoring) are particularly sensitive to this constraint. LLMs offer a natural way out: personalized exchanges at zero marginal cost and uniform quality. We run a nationwide experiment in Chile ($N \approx 42,000$) designed to reduce teacher shortages by increasing applications to education majors. We compare an AI chatbot, *Kai*, with trained human counselors, both reaching high school seniors via WhatsApp. Among students with baseline interest in education (the intervention’s target), Kai significantly increases first-ranked education majors and the share of education majors in students’ choice sets; estimates in the human arm are smaller and not significant. Although the best human counselors are highly effective, Kai matches or outperforms roughly two-thirds of them. Its main advantage lies in what has been called “controlled variance”: it compresses the dispersion in quality observed across human implementers. Finally, text analysis reveals different interaction patterns across arms. Kai is notably faster (and thus conversations are more dynamic), it concentrates on factual content and scores higher on semantic coherence, while counselors score substantially higher on motivational and empathetic language.

Keywords: teachers, teacher shortages, scale-up, behavioral interventions, AI chatbots, generative AI.

JEL Classification: D91, I23, I25.

*We thank the Elige Educar team for their collaboration in this study. This research was approved by the HML Institutional Review Board (Study #2713). RCT registry: AEARCTR-0014942. A previous version of this paper circulated under the title *Can AI Replace Human Counselors at Scale? A Nationwide Experiment to Reduce Teacher Shortages*. All errors are our own.

[†]McGill, CEPR, J-PAL. E-mail: nicolas.ajzenman@mcgill.ca.

[‡]Inter-American Development Bank. E-mail: anade@iadb.org.

[§]Inter-American Development Bank. E-mail: gregorye@iadb.org.

[¶]E-mail: calbornozh@fen.uchile.cl.

^{||}Boston University. E-mail: sveleda@bu.edu.

1 Introduction

Decades of experimental evidence show that well-designed behavioral interventions can meaningfully shift individual decisions, yet their effectiveness depends critically on the quality of those who implement them (Davis et al., 2017; Vivalt, 2020). High-quality implementers are scarce: as programs scale, the implementer pool broadens and average quality falls (Al-Ubaydli et al., 2017b, 2020), a key explanation for why promising programs often fail to replicate at scale (Kraft et al., 2018).

A large share of high-impact behavioral interventions is conversation-based (counseling sessions, peer mentoring, and direct messaging platforms), which makes them particularly sensitive to implementer quality and therefore especially vulnerable to fade-out at scale (Crépon et al., 2013; Blattman et al., 2017; Carrell and Sacerdote, 2017). Generative AI offers a potential solution when such interventions are delivered through digital platforms, an increasingly common mode of delivery (Ajzenman et al., 2026). Large language models can deliver personalized conversations of uniform quality, at almost zero marginal cost, eliminating the variation in implementer quality that emerges as human-staffed programs expand. Whether AI can substitute for the interpersonal qualities thought to underpin the effectiveness of these interventions, however, is an open empirical question. We address this question using a nationwide field experiment in Chile designed to motivate high school seniors toward teaching careers, comparing two delivery arms: one staffed by trained human counselors, the other by an AI-powered chatbot.

We find that the AI chatbot is at least as effective as trained human counselors at shifting students’ college-application choices (at a fraction of the cost), providing evidence that AI can relax the implementer-quality constraint that limits the scale-up of conversation-based interventions. Moreover, beyond the fact that AI’s average performance seems to be at least as good as that of the average human counselor, a great advantage is that it homogenizes the quality of delivery across interactions, eliminating the quality heterogeneity observed among human implementers.

The program lasted 10 weeks and contacted more than 41,000 students. Students in the human arm chatted on WhatsApp with trained counselors. Students in the AI arm received the same content from Kai, a WhatsApp chatbot built on GPT-4o mini that can hold open-ended, personalized conversations. Kai was trained on the materials used to prepare the human counselors, and its prompts were adjusted over several rounds of pilot testing until its style was close to theirs. The scripts and training material draw on behavioral insights emphasizing motivational factors (intrinsic and extrinsic), alongside objective information about the college application process and teaching careers in Chile. Elige Educar, a Chilean NGO co-leading the implementation, designed the scripts, content, and style to be as similar as possible across arms. The intervention was timed to coincide with the Chilean college application cycle. In Chile, university admission is centralized. Students submit a ranked list of up to ten university-

major combinations, and admission depends on a weighted average of their high school GPA and their scores on the national entrance exam (PAES). The treatment took place in the weeks just before students submitted these lists. Students in both arms were recruited in the same way and consented to be contacted without knowing which arm they were in, although the first message told them whether they would be talking to a human counselor or to an AI chatbot.

We document several results. Among students with baseline interest in teaching – the program’s intended target – those assigned to the AI chatbot arm were 1.4 percentage points more likely to list an education major as their first choice (control mean: 26.1%) and had a 1.1 pp higher share of education majors in their choice set (control mean: 24.4%), both significant at the 5% level. We find no effect among students without baseline interest in education, consistent with the program’s targeting logic. The human counseling arm produced smaller and statistically insignificant effects. Among students interested in education careers, the human arm increased the probability of listing an education major as their first choice by 1.1 pp (insignificant at conventional levels). The human arm’s effect on the share of education majors in the choice set was similarly insignificant, with a point estimate roughly half that of the AI chatbot. The chatbot arm had a larger sample size, which raises the question of whether the human arm’s null result reflects limited power rather than a true zero effect. Although possible, this seems unlikely to drive the null: given the human arm’s very large sample size (4,670 effectively treated students plus 7,781 controls, considering only those interested in education), the design was powered to detect even small effects at conventional significance levels (including the point estimate of the kai effect). In any case, a cautious interpretation would be that the chatbot performed at least as well as the human arm.¹

Interestingly, the average performance of humans and Kai masks substantial heterogeneity. As Figure 1 shows, the best human counselors are highly effective: the proportion of students who ranked an education major first is similar to or higher than under Kai for more than a third of the counselors. Kai’s average performance is above that of the average human counselor, but its main advantage is that it provides what [Jabarian and Henkel \(2026\)](#) call “controlled variance,” homogenizing quality across interactions and thus eliminating the wide dispersion observed among human implementers.

We then turn to mechanisms. Although for privacy reasons we cannot link individual conversations to student-level outcomes, we analyze the content and dynamics of student-counselor exchanges in each arm to understand how AI and human counselors differ in engagement. Drawing on recent methodological advances in the analysis of conversational transcripts ([Jabarian and Henkel, 2026](#)), we use text-analysis techniques to study several features of these exchanges.

¹Effects on enrollment in an education program are not significant for the chatbot arm and marginally significant for the human arm (and insignificant after correcting for multiple hypotheses). We interpret these results with special caution, since SUTVA does not hold for this outcome: treated students’ decisions affect control students’ allocation through the centralized assignment mechanism, which precludes a causal interpretation of treatment effects on enrollment (see [Karnani, 2024](#)).

We begin our analysis by showing that, conditional on a conversation starting, conversations with human counselors contain a greater total number of messages than conversations with the chatbot (8 versus 7), a pattern that admits two competing interpretations: longer exchanges may reflect deeper engagement with human counselors, or, alternatively, less efficient information delivery, with the chatbot resolving the student’s query in fewer turns rather than the student disengaging. To distinguish between them, we compare the *breadth* of conversations across arms: despite exchanging substantially fewer messages per user, the chatbot covers all 18 subtopics that human counselors address and matches or exceeds human coverage in 11 of them, which we interpret as evidence that shorter chatbot conversations reflect more efficient information delivery rather than disengagement.

Although the thematic coverage is similar in breadth, the topics themselves differ substantially in nature. Conversations with the AI chatbot concentrate on admissions processes, entrance requirements, scholarships, and university rankings, while those with human counselors gravitate toward vocational guidance, personal experiences, and open-ended conversational exchanges. A more detailed classification, using a Chilean Spanish BERT model fine-tuned on a manually labeled sample of messages classified as factual or subjective, confirms this pattern: the share of exchanges classified as factual is significantly larger for the chatbot (25% versus 15%), while the reverse holds for subjective exchanges (4% versus 11%).

A further dimension along which conversations might differ is their pace and dynamism. Whether a student stays in the conversation probably depends on how well the counselor holds their attention, and here the chatbot has some mechanical advantages. It is available at any hour. Human counselors work in shifts, and each of them handles many more students (there were 45 counselors, while the chatbot’s capacity is effectively unlimited). Even when a human counselor is available, the chatbot responds faster: it does not need to compose, type, or look up information. Consistent with this, the chatbot dominates on every dynamism metric we examine. Median conversation duration per student was 1.12 minutes with Kai and 28.23 minutes with human counselors; response times are near-instantaneous for the chatbot versus roughly one minute for counselors; the median gap between consecutive student messages is about one minute versus roughly three; and time to the first informative exchange is under 44 seconds versus approximately three minutes. Since both arms cover a similar range of topics, we interpret these gaps as differences in the *dynamism* of the exchange, with the chatbot delivering similar content in less time and with quicker replies. These dynamism advantages need not translate into better counseling – slower human responses may reflect deliberation – a possibility we examine using direct measures of conversational quality below.

We next examine whether conversational quality differs across arms, independently of the topics discussed or the speed at which they unfold, using several measures grouped into two categories: informational coherence and interpersonal quality. We compute the cosine similarity between dense representations of each prompt and its reply, in both directions (student-

to-counselor and counselor-to-student), as a measure of how tightly each response tracks the question that elicited it. In the literature on dialogue-evaluation, embedding-based alignment is a standard proxy for response quality (Tao et al., 2018; Mehri and Eskenazi, 2020), and in counseling contexts it correlates with the quality of counselor-client exchange (Pérez-Rosas et al., 2019). Kai-led conversations exhibit higher coherence than human-led ones in both directions: Kai provides more semantically relevant responses to students’ questions, and students in turn respond more coherently to Kai’s prompts than to those of human counselors.

We then turn to the interpersonal dimension of quality, focusing on *empathy*, *motivational support* – attributes where human counselors are expected to hold a comparative advantage – and *informativeness*, which captures the clarity, relevance, and actionability of the information provided. A team of trained raters (Chilean psychology students) scored a random sample of conversations on these dimensions on a 1–3 scale, with an additional “not applicable” category for cases in which the conversation did not provide a meaningful opportunity for the counselor to display the corresponding skill, and we use a hybrid classification model to scale the ratings to the full sample (see Appendix A). Human counselors substantially outperform Kai on *motivation* and *empathy*, while performing similarly or slightly worse on *informativeness*. This pattern is consistent with the content differences documented above: Kai’s replies are more factual and precise, whereas human counselors rely more heavily on socioemotional content and come across as warmer and more encouraging.

Taken together, these patterns reveal a clear tradeoff in the nature of human and AI-led counseling. The AI chatbot delivers information more efficiently, covering a similar range of topics – with greater emphasis on factual content – with higher conversational coherence and substantially more dynamic interactions, but falls short on the interpersonal qualities that distinguish human counselors.

These conversational attributes exhibit substantial heterogeneity across human counselors. Even on the dimensions where Kai outperforms the average human counselor, such as *topic breadth*, *subject coherence*, or rater-scored *informativeness*, a substantial share of counselors at the top of the distribution match or surpass the chatbot. Conversely, on the dimensions where human counselors hold a clear average advantage, such as *motivation* and *empathy*, a small fraction at the bottom of the distribution fails to outperform Kai. This heterogeneity highlights a well-known challenge: maintaining large treatment effects at scale (Al-Ubaydli et al., 2017b, 2020). The top counselors match (or even outperform) Kai, but maintaining that level of quality across the full workforce becomes increasingly difficult as the program grows.

Our paper contributes to several strands of the literature. First, we contribute to a growing field-experimental literature on AI performance in complex tasks. While one strand shows that AI assistance raises the productivity of human workers (Brynjolfsson et al., 2025; Noy and Zhang, 2023; Dell’Acqua et al., 2026; Cui et al., 2025), a second, less developed strand deploys AI to substitute for humans in structured interactions such as qualitative interviews,

recruiting, motivational interviewing, and medical or mental-health consultations (Geiecke and Jaravel, 2024; Jabarian and Henkel, 2026; Chopra et al., 2026; Chen et al., 2025; Heinz et al., 2025). We contribute to this second strand in two dimensions: ours is a nationwide intervention targeting a consequential, long-run decision (whether to pursue a teaching career), and rather than reporting an aggregate AI-human comparison, we decompose the interaction into conversational dimensions, showing *why* autonomous AI can be at least as effective as humans in terms of performance.

Second, we contribute to the growing literature on scaling-up experimental interventions. A central finding of this literature is that the effects of the intervention are attenuated at scale, in part because the pool of available implementers expands and average quality falls (Al-Ubaydli et al., 2017b, 2020, 2021; Muralidharan and Niehaus, 2017; Vivalt, 2020; DellaVigna and Linos, 2022; Ajzenman et al., 2026). We contribute to this literature in two ways. First: we provide direct experimental evidence that the implementer-quality channel binds in practice: cross-counselor variation in conversational quality in our setting is large enough that an AI chatbot delivering uniform quality outperforms the average human counselor in some dimensions; and we go beyond aggregate comparisons and document the conversational attributes along which the two arms differ: AI leads on informational coherence and consistency, while humans retain a clear advantage in interpersonal warmth.

Third, we contribute to the literature on teacher supply (Corcoran et al., 2004; Hoxby and Leigh, 2004; Boyd et al., 2005, 2012; Bacolod, 2007) and the motivations underlying career choices in education (Han et al., 2018; Perryman and Calvert, 2020). While most of the existing work has focused on monetary incentives—such as salary increases—as a tool to attract and retain teachers (e.g., Clotfelter et al., 2008; Falch, 2011; de Ree et al., 2018); growing evidence indicates that students’ decisions to enter teaching are also shaped by other motives (for instance, intrinsic, extrinsic, and prestige-related motivations) (Han et al., 2018; Perryman and Calvert, 2020). Yet causal evidence on whether interventions that target these motivations can effectively shift teacher supply remains limited (Ajzenman et al., 2021; Christian et al., 2024; Ajzenman et al., 2026). We extend this literature by providing the first experimental evidence that AI-based behaviorally informed counseling delivery can shift career choice.

Finally, we contribute to the emerging literature on automation versus augmentation in AI (Acemoglu et al., 2026; Acemoglu, 2025). Recent theoretical work predicts that LLMs will substitute for workers in routine tasks that can be codified and complement them in tasks that require judgment (Ide and Talamàs, 2024; Korinek and Suh, 2024). In our experiment, an AI chatbot fully replaced human counselors in a non-routine task that relies heavily on personal interaction, and it performed at least as well (a case of *full automation*). A plausible interpretation could be that the relational intensity of a task matters for the line between substitution and augmentation only to the extent that the relational dimension affects outcomes. We can test this directly, because we observe both interpersonal quality measures and treatment

effects.

Our findings have direct policy implications. Scaling up interventions without degrading delivery quality is challenging: maintaining quality typically raises per-unit costs, and in some cases high-quality human capital is too scarce to scale without losses in implementation fidelity (Davis et al., 2017; Al-Ubaydli et al., 2017a). Our results show that, when implementer quality is heterogeneous, AI delivery can match or outperform the average human implementer at a substantially lower cost. The human arm cost approximately \$6.20 per student contacted (counselor remuneration, coordination, and messaging), implying roughly \$223 per additional student induced to rank an education major first. Kai cost approximately \$1.95 per student contacted (including development costs amortized over five years), or \$123 per additional student induced (taking point estimates at face value; since the human-arm effect is statistically insignificant): roughly twice as cost-effective, and without any binding capacity constraint on scale. Excluding development costs, these figures fall to \$0.95 and \$60, roughly four times the cost-effectiveness of the human arm. Given the pace of progress in AI technology, these costs are likely to continue to fall. The gap should only widen.

2 Context

2.1 Teacher shortages in Chile

Teachers are a crucial input in the educational production function, representing the most significant in-school factor affecting student learning and often accounting for 60-80% of national education budgets (Rockoff, 2004; Rivkin et al., 2005; Staiger and Rockoff, 2010; Chetty et al., 2011, 2014). However, attracting capable candidates into teaching is a persistent challenge (Elacqua et al., 2020). While many countries have introduced policies to attract good teachers to the regions where they are most needed (Clotfelter et al., 2008; Jackson, 2009; Ajzenman et al., 2021, 2023; Elacqua et al., 2022), interventions that target the entry decision itself are less common (Ajzenman et al., 2021, 2026).

Chile faces a growing and severe shortage of qualified teachers. The Ministry of Education’s most recent national assessment (Ministerio de Educación, Centro de Estudios and Elige Educar, 2025) estimates that schools are short about 123,789 weekly hours taught by qualified teachers (*docentes idóneos*), or 6.2% of the hours required. If only subject-specialist teachers are counted, the shortfall rises to 235,602 hours (11.7% of national need). Rural schools and schools serving the most socioeconomically vulnerable students have the largest deficits. In rural schools, 8.3% of hours lack a qualified teacher and 13.2% lack a specialist, both well above the national figures. By subject, the largest gap is in Technological Education, where more than 90% of secondary schools have no specialist teacher, followed by Religion and Philosophy, English, and the Sciences (Ministerio de Educación, Centro de Estudios and Elige Educar,

2025).

Chile has passed several reforms over the past decade to strengthen the teaching profession. The *Beca Vocación de Profesor* (BVP) is a scholarship for high-performing applicants to teaching programs (Pérez-Núñez, 2020). Free postsecondary tuition was extended to students in the bottom 60% of the income distribution. A teacher career law (Ley 20.903), phased in from 2017, raised entry requirements for teacher education programs, increased teacher salaries, reduced instructional hours, and introduced structured induction support (Ávalos and Bellei, 2019). The shortage persists. Some reforms even reduced entry into teaching for a while. Once tuition was free, high-achieving low-income students could afford majors with higher returns, and fewer of them chose teaching (Castro-Zarzur et al., 2022). Total enrollment in teacher education also fell in the short run after Ley 20.903 raised entry requirements (UNDP, 2023). Structural reforms alone are unlikely to close the gap. Cheaper interventions that target motivational and informational barriers at the time of career choice could complement them.

Behavioral and informational frictions matter too. According to Elige Educar, CIAE y Universidad de Magallanes (2023), a sizable share of students who express interest in education do not apply to teaching programs in the end. They cite perceived low salaries, worries about prestige, and lack of information about working conditions and careers. Causal evidence on whether targeted, low-cost interventions can shift these decisions at scale remains limited (Ajzenman et al., 2026; Christian et al., 2024).

2.2 The Chilean university admissions system

Chile’s higher education system comprises universities, professional institutes, and technical-vocational schools. Only certified universities may offer teacher education programs. Students apply through a nationwide centralized admissions system that allocates them to university-major combinations based on their stated preference rankings and their academic performance on the national university entrance exam (PAES, *Prueba de Acceso a Educación Superior*), administered by DEMRE, the higher education testing arm of the Universidad de Chile.

The 2025 admissions cycle ran from June 2024 (PAES registration) to January 2025 (program assignment). Registration was open from June to July; the test was administered in the first week of December; results were released in January 2025. Following the release of the results, the application period opened, allowing students to rank up to ten university-major combinations in order of preference. Students were then matched to programs via the student-proposing deferred acceptance algorithm (Gale and Shapley, 1962), with admission results published at the end of January 2025. Programs rank applicants using a weighted score combining mathematics, and verbal test scores, high school GPA (NEM), and a social-science or science test score, with weights varying by program; teacher education programs require a minimum PAES score of 502.

When registering for the PAES, students complete a survey administered by DEMRE that collects information on their academic interests and preferences. One question asks them to identify their vocational interest as Education, other social sciences, or none of these (more concretely, they could choose from a list of ten fields). Although the intervention primarily targets students who selected Education, our sample also includes students who selected Social Sciences or Humanities but not Education. The survey also includes a consent question asking students whether they agree to have their contact information shared with Elige Educar for counseling assistance during the higher education application process. The study protocol, including the consent language and the random-assignment design, was approved by an Institutional Review Board (IRB) prior to implementation. Students could withdraw at any time during the intervention via WhatsApp. A post-intervention debriefing offered a final opportunity to withdraw from data use. Students who consented (and did not withdraw) form the sampling frame from which the experimental sample was drawn for this study. The same system also produces the administrative application and enrollment records used as the primary outcome data; see Section 3.

2.3 The Quiero Ser Profe program

Elige Educar is a Chilean NGO founded in 2009 with the mission of ensuring that all children have access to quality teachers. Its flagship teacher recruitment initiative, the *Quiero Ser Profe* program (QSP) has operated in partnership with DEMRE and the Chilean Ministry of Education since 2010.

The QSP program addresses the barriers documented in the previous section – perceived low salaries, prestige concerns, and information gaps about the profession and the application process – through three mutually reinforcing components. First, a large-scale annual media campaign promotes a positive image of the teaching profession through television, radio, and social media. Second, the Elige Educar team provides information at career fairs at universities across Chile. Third and most relevant for this study, individualized WhatsApp counseling connects students interested in education careers with trained counselors, typically advanced education students who have recently chosen teaching, making peer modeling an intentional design feature.

Counseling conversations serve two complementary purposes: providing clear and timely responses to students’ most common questions – admission requirements, scholarships, labor market conditions, and teacher salaries, and program-level information on accreditation and curricula – and engaging students in a reflective exploration of whether teaching is the right path for them. Counselors are trained to tailor each conversation to the student’s individual concerns and motivational profile.

Between 2018 and 2020, QSP contacted approximately 2,400–3,300 students per year

through personalized WhatsApp counseling, with a team of 30–45 counselors and an annual budget of approximately 133 million Chilean pesos. A 2020 internal evaluation found that 63 percent of participants reported intending to enroll in a teacher education program (up from 55 percent in 2019), and that one-third of initially uncertain students ultimately chose a teaching path. These figures, however, came from self-reported survey responses in a non-representative sample and did not establish a causal effect of the program (Elige Educar, 2025).

This gap motivated a collaboration between Elige Educar and the Inter-American Development Bank (IDB) to evaluate the program through a randomized controlled trial. The first evaluation, in 2021, compared human counseling with a rule-based chatbot, a menu-driven WhatsApp tool that offered the same information as the counselors through a predefined decision tree (Ajzenman et al., 2026). Human counselors increased the probability that a student ranked a teaching program first by 1.25 percentage points. The chatbot’s effects were about a third as large and statistically insignificant, a difference attributed in part to its limited ability to personalize conversations.

This paper evaluates the second phase of this scaling strategy, implemented in 2024. Human counseling can shift students’ career choices, but the program’s capacity is limited. A pool of 45 counselors can serve about 9,000 students per admissions cycle, less than half of the students who express interest in education careers each year. To overcome this constraint while preserving the program’s personalized counseling format, Elige Educar and the IDB developed Kai, a WhatsApp-based generative AI chatbot, as a candidate to fully automate the counseling function. Kai, unlike the earlier rule-based chatbot, can hold open-ended and personalized conversations. It uses the same knowledge base and motivational content as the training of human counselors, and it can serve an effectively unlimited number of students at a low marginal cost. The experiment tests whether a generative AI system can match or exceed the effects of trained human counselors, and examines the mechanisms behind any difference.

3 Experimental Design

3.1 The *Quiero Ser Profe* Human Counseling Arm

Students randomly assigned to the human counseling arm received personalized guidance via WhatsApp from trained human counselors following the standard QSP methodology. The human arm was coordinated by two program coordinators from the Elige Educar team who were responsible for counselor recruitment, training, operational oversight, and day-to-day support throughout the intervention period. The pool comprised 45 counselors recruited and trained by Elige Educar.

The counselor pool was predominantly female (75.6%) and composed largely of advanced education students: 35.6% were in their fourth year of a teaching degree, 28.9% in their fifth

year, and 17.8% in their third year, with the remaining 17.7% either recently graduated or working as teachers. About half (53.3%) had participated in the program in a previous cycle. Counselors came from several universities, most often Universidad de Concepción in the Bío-Bío Region (36%) and Universidad Alberto Hurtado in the Metropolitan Region (18%), with the rest at other institutions across the country. They specialized in language and literature, history, mathematics, early childhood education, the arts, and other teaching fields.

Counselor training followed the QSP methodology. Under this methodology, guidance is structured but adapted to each student, and counselors draw on their own experience as education students to orient them. Counselors are expected to give clear and accessible information on scholarships, admissions, and programs, and to make students comfortable enough to voice their doubts and ask questions. They are also asked to be warm, patient, and clear, and to take a genuine interest in each student’s situation. Training covered both what counselors should discuss and how they should communicate. On the content side, training addressed the topics most commonly arising in student interactions: university programs, accreditation standards, entry requirements, scholarship and financing options (including the Beca Vocación de Profesor, FUAS, free tuition), the centralized admissions process, and labor market conditions for teachers including salaries, employability, and career development pathways. The communicative approach trained counselors to identify students’ motivational profiles at the outset of a conversation and to tailor their engagement accordingly, following the approach developed in [Ajzenman et al. \(2021\)](#) and in subsequent iterations of the program. Counselors were also trained to handle emotionally sensitive conversations (including students facing discouraging family expectations or significant vocational uncertainty) and were encouraged to draw on their own experiences as education students as a form of peer modeling.

Each counselor was assigned approximately 23 students per week and initiated contact using standardized opening templates and a semi-structured conversation guide (*guión*). The guión organized counseling interactions into five sequential blocks (counselor introduction, student contextualization, motivational engagement, information provision, and closing), but was designed to be flexible: counselors were expected to adapt the flow, tone, and emphasis of each conversation to the student’s expressed needs, concerns, and motivational profile. Within each block, the guide provided example texts and suggested talking points, but counselors retained discretion over wording. Counselors operated in scheduled afternoon and evening shifts, prioritizing active conversations; each conversation window remained open for 24 hours after the student’s last message.

3.2 The *Quiero Ser Profe* Kai Chatbot Counseling Arm

Students randomly assigned to the chatbot arm received personalized guidance via WhatsApp from Kai, an AI counselor designed as a direct counterpart to the human arm. Kai was built to

mirror the human arm’s content and conversational style as closely as possible while delivering personalized counseling at unlimited scale.

Kai was developed over a four-month period from July through October 2024. Its development was led by Patagonian, a specialized IT firm, in collaboration with the Elige Educar team. Kai is a WhatsApp-based conversational agent built on GPT-4o-mini and designed to engage students in an open-ended, personalized dialogue about teaching careers and the university application process. The rule-based chatbot evaluated in [Ajzenman et al. \(2026\)](#) followed a predefined decision tree. Kai, a generative AI system, reads students’ free-text messages and writes its own replies.

Kai drew on a content bank that compiled the QSP program’s methodology and expertise. It searched this bank through a Retrieval-Augmented Generation (RAG) architecture with semantic search, so it could find the relevant material even when a student’s wording differed from the terms used in the program documents. The bank had three components. The first contained information on the topics that come up most often in counseling conversations. These include university programs, accreditation, and score requirements; financing (the *Beca Vocación de Profesor*, FUAS, and free tuition); admissions requirements and procedures; and labor market prospects such as teacher salaries, employability, and professional development. The second covered the socioemotional and vocational side of the decision, such as students’ motivational profiles, their values and identity, their life plans, and how they deal with their own, their family’s, and society’s expectations about becoming a teacher. Elige Educar’s expert counselors worked closely on the third component, written for the most frequent and the most difficult queries. It covered common questions about programs, financing, and universities. It also covered harder conversations, such as how to support a student whose family discouraged a teaching career, concerns about teachers’ working conditions and social recognition, and doubts about fit with the profession.

Like the counselors, Kai worked through WhatsApp, so the channel is the same in both arms. This matters for interpretation: because both arms use the same medium, any difference in outcomes between them reflects the counselor (human or AI), not the platform. It was available 24 hours a day.

Elige Educar is committed to inclusive communication, and students’ gender was unknown at first contact, so Kai’s persona and name were gender-neutral. It wrote in a warm, accessible and direct Chilean Spanish register suited to the students. Responses had a 100-word cap. In earlier cycles, the counselor messages that worked best were short. Kai never claimed to be human. The first message said that Kai was an AI counselor. When a student asked whether they were talking to a person, Kai repeated that it was an AI assistant trained on the QSP methodology.

QSP counselors were encouraged to share their own story as education students (why they chose teaching, what was hard, what was rewarding) as a form of social modeling. Kai had

no such story, and making one up would have broken the no-deception design. The design team expected Kai’s consistency, availability, and speed to make up for this. Earlier evidence from the program also suggested that information, more than anything else, drove changes in students’ stated preferences.

Reviewers from the IDB and Elige Educar, two of them experienced QSP counselors, tested Kai for several weeks before launch. In the first rounds they checked technical functioning and content: whether Kai answered the expected student questions accurately and to the point, whether its information on admission requirements, scholarships, and university programs was correct and up to date, and whether its tone was warm, accessible, and close to the human counselors’. Later rounds focused on ethically sensitive or emotional exchanges. Kai had to respond with care to students in distress or great uncertainty without going into matters that need professional support, and to decline questions outside its scope and refer the student elsewhere instead of guessing. Each round ended with structured feedback, and the developers adjusted Kai’s prompt, knowledge base, and response parameters before the next one.

3.3 Intervention Design and Contact Protocol

Both arms used Elige Educar’s WhatsApp Business account, so students in each arm saw the same platform, sender, and interface. The first message said who the sender was. Students in the chatbot arm were greeted by Kai, introduced as an AI-powered counselor, and students in the human arm received an equivalent message from their assigned counselor. The message had four buttons. Three of them started the conversation in different ways (“How do I work?”, “How can I help you?”, and “Let’s get started!”). The fourth (“I don’t want to participate”) closed the student’s file at once, and the student was not contacted again. Students who stayed in the program could write to their counselor or to Kai whenever they wanted during the intervention.

The contact protocol was the same in both arms. Students were split into eight groups, and each group was contacted in a given week, starting on October 28, 2024. A student who did not answer the first message that week received a second one, and no more. Anyone who chose the opt-out option, in any message, was not contacted again.

The intervention had two phases. The first covered the eight contact waves, from late October to early December 2024. Its aim was to motivate students to consider education careers and to give them information for their decision before the application period. In the week of January 6, 2025, all students who had not opted out were contacted again, this time with support for the university application itself. Students could keep starting conversations until the intervention ended on January 14, 2025.

3.4 Sample Selection and Randomization

We identified 151,253 students who met the eligibility criteria for the study. These were students registered to take the PAES who, when completing the DEMRE registration survey in July 2024, declared an interest in pursuing a career in education or social sciences and consented to share their information and be contacted by Elige Educar for application counseling and research. From this pool we drew a sample of 43,539 students, of whom 23,690 had declared an interest in education careers and 19,849 an interest in social sciences but not education. Students interested in education are the program’s main target. We include the social sciences group to test whether the intervention can also change the career preferences of students who had not expressed interest in teaching. In an earlier evaluation of the program, [Ajzenman et al. \(2026\)](#) found no effects for this group, which gives us a benchmark.

We randomly assigned students to three arms: human counseling (8,998 students), the chatbot (20,000 students), and a passive control group that received no intervention (14,539 students). The arms differ in size because of Elige Educar’s capacity. Its 45 counselors could serve at most 9,000 students during the intervention, while the chatbot arm had no such limit, so we made it large to maximize statistical precision. We stratified the randomization by gender, type of high school attended (public, subsidized private, or private), and declared career interest (education or social sciences). Students left in incomplete strata because of small cell sizes (“misfits”) were randomized as a separate stratum.

The final estimation sample excludes students who did not sit the PAES (361 observations) and those who opted out of the program via any contact message (1,743 observations), yielding a final sample of 41,432 students, of whom 22,739 declared interest in education and 18,693 in social sciences.

Finally, as stated in the pre-registration, we estimate effects separately for two groups defined at baseline. The DEMRE survey asks students, “In which area would you like to study or work?”, and lists ten areas (Education, Humanities, Arts and Architecture, Health, Technology, Social Sciences, Agriculture, Basic Sciences, Business and Commerce, and Law), each accompanied by examples of specific majors to clarify its scope. We classify a student as *interested in Education* if they marked Education, regardless of whether they also marked other areas, and as interested in social sciences if they marked Social Sciences or Humanities, but not Education. For brevity, we refer to this group as ‘social sciences’ throughout the paper; it includes students who selected Humanities.

4 Data

4.1 Administrative Data on University Applications

We use anonymized administrative data from DEMRE, the Chilean agency responsible for overseeing the centralized university admissions process. We linked these records to the experimental sample through masked student identifiers assigned by Elige Educar. Researchers never had access to personally identifying information.

For each student, the data record university applications, enrollment, and a set of baseline covariates collected at PAES registration. We use four outcome variables. (i) *Ranked first* equals one if the student listed an education major as their first-choice university-program combination. (ii) *Proportion listed* is the share of education majors on the student’s list, which can include up to ten programs. (iii) For *At least one listed*, it is enough that one program anywhere on the list is an education major. (iv) *Enrolled* is one for students who ended up in an education program, whether they entered through the PAES or through a special admissions pathway. The first three outcomes describe the direction and intensity of the preferences students stated before the centralized matching, and the fourth is realized enrollment. Enrollment has to be read with caution. Under the centralized matching algorithm (Gale-Shapley), the choices of treated students change where other students are placed, which violates SUTVA (see [Karnani \(2024\)](#)).

All baseline covariates come from PAES registration. Schools are fully public (state-owned and publicly funded), subsidized private (privately owned, with a public subsidy per student), or private (funded only by tuition). The NEM score is the student’s high school GPA converted to the 0–1,000 PAES scale. The ranking score uses the same scale but compares a student’s GPA with those of their cohort and the previous three graduating classes at the same school. Other covariates are gender, age in 2025, the mean of the two mandatory PAES tests (reading comprehension and mathematics), self-reported Chilean nationality and indigenous ethnicity, and whether the student lives in the Metropolitan Region. ODEPA classifies municipalities into three rurality levels by population density, using the 2024 census. The family income decile comes from the government’s household vulnerability records, which mix self-reported and administrative data. Missing values are set to zero, and each specification includes missing-value indicators.

4.2 WhatsApp Interaction Data

We also have every WhatsApp message exchanged between students and their counselor (human or AI) during the intervention, with its sender and time stamp. The conversations are anonymized and linked to the experimental sample through the same masked identifiers as the administrative data, and the analysis dataset contains no personally identifying information.

The data-use agreement lets us link conversations to some pre-treatment variables. It does not let us link them to outcomes.

The mechanisms analysis in Section 7 relies on these data. We measure conversation length, the timing and order of messages, topics, whether content is factual or subjective, and how semantically coherent each reply is with the prompt before it. Appendix A gives the details of each measure.

Chat users often break one idea into several short messages a few seconds apart, so before the analysis we merge consecutive messages from the same participant sent within 30 seconds into one “message burst,” which approximates a full conversational turn.

After applying the sample restrictions described in Appendix A, retaining students who declared an interest in education careers at PAES registration, sat the PAES, and actively interacted with their counselor, and excluding those who opted out, the analytic sample comprises 36,860 messages in the human arm and 37,562 messages in the Kai arm.

4.3 Descriptive Statistics and Balance Checks

The estimation sample of 41,432 students is 65.0% female. The majority attend subsidized private schools (57.2%), which are privately owned but receive per-student public subsidies; 35.9% attend fully public schools, and the remaining 6.9% attend private nonsubsidized schools. About a quarter (25.9%) attend technical-vocational schools. Just over half (54.9%) declared an interest in education careers at PAES registration, and the rest in social sciences. Family income falls on average in the fourth decile of the national distribution. In terms of baseline outcomes, 15.5% of students ranked an education major first in their choice set, 25.8% included at least one education major, and 13.6% ultimately enrolled in an education program. Table 1 presents full descriptive statistics for all covariates and outcomes.

Table 2 shows that the sample is balanced across nearly every observed dimension. The few differences that emerge are small in magnitude. Omnibus F -tests of joint orthogonality using randomization inference (Kerwin et al., 2024) confirm balance across all pairwise comparisons. Reassuringly, Tables 3 and 4 show that this pattern of balance broadly holds within each interest subgroup.

5 Empirical Strategy

To estimate the effect of each arm of the intervention on students’ educational choices, we run the following specification:

$$y_i = \alpha_1 \text{Human}_i + \alpha_2 \text{Kai}_i + X_i' \beta + \varepsilon_i \quad (1)$$

where y_i is one of the four outcome variables described in Section 4 for student i ; Human_i and Kai_i are indicators for assignment to the human counseling arm and the Kai chatbot arm, respectively, with the passive control group as the omitted category; and X_i includes GPA score (NEM), ranking score, average math and verbal test score, age, and indicators for attending a technical-vocational school, holding Chilean nationality, and belonging to an indigenous people, as well as fixed effects for randomization strata (gender, school type, and interest in education versus social sciences), family income decile, Metropolitan Region, rurality of the municipality of residence, and missing value indicators for any control variable. The coefficients of interest are α_1 and α_2 , which capture the intention-to-treat (ITT) effects of each arm relative to the control group. We also report a direct test of $\alpha_1 = \alpha_2$ to assess whether the two arms differ from each other.

Following our pre-analysis plan, we estimate equation (1) on three samples: the subsample of students who, prior to the intervention, declared an interest in education; the subsample of those who did not; and the full sample combining both groups. We pre-registered this split for two reasons. First, in the earlier evaluation (Ajzenman et al., 2026), effects were muted for students without prior interest in education and concentrated among those with it. Second, the program is meant for students who are already considering teaching. We still include students interested in social sciences, as a benchmark: can the intervention pull students from a neighboring field? Ajzenman et al. (2026) suggest it cannot. See Section 3.4 for details on the survey question used to construct this classification from baseline data.

Since randomization was conducted at the individual level, all specifications report heteroskedasticity-robust standard errors. As a robustness check for multiple testing, we report Benjamini-Hochberg adjusted p -values for the family of primary outcomes.

6 Results

Table 5 reports the intent-to-treat effects of each arm on our four application and enrollment outcomes, estimated from equation (1). The three panels correspond to the three estimation samples: Panel A restricts attention to students who at PAES registration declared an interest in education, Panel B to those who declared an interest in social sciences but not education, and Panel C pools both groups.

We start with Panel A, the target sample for the *Quiero Ser Profe* program. We present two types of outcomes: *Choice* outcomes, which reflect students’ preferences over majors and the priorities they assign to them; and *enrollment*, which requires a caveat before interpretation. Choice outcomes depend only on the student’s own behavior. Enrollment is decided by a centralized allocation mechanism, where each student’s assignment depends on what every other applicant chose. This has two consequences. First, enrollment is a general-equilibrium

variable. Whether a treated student is admitted also depends on program capacity and on the choices of peers. Second, and more consequentially for inference, enrollment outcomes are subject to between-arm spillover effects that violate SUTVA: students’ choices in any arm affect other students’ enrollment in any other arm (see [Karnani, 2024](#)). We thus interpret the enrollment outcome with caution, since a causal reading is not clean.

[Ajzenman et al. \(2026\)](#) provide a benchmark for the results that follow: they evaluated a structurally similar version of the program in 2021, with minor implementation changes. The human arm produced effects of 1.25 and 1.28 percentage points on *Ranked First* and *Proportion*, respectively, and a null effect on the extensive-margin outcome *Listed at Least Once*.

Table 5 Panel A (students with prior interest in education), shows that treatment with Kai increases the probability that a student lists an education program as their first choice (*Ranked First*) by 1.4 percentage points, significant at the 5% level (with a control mean of 26.1%). The human arm produces a point estimate of 1.1 percentage points, insignificant at conventional levels. Kai also delivers a positive effect of 1.1 percentage points on the share of education programs within the student’s choice set (*Proportion*; over a control mean of 24.4%), significant at the 5% level. The human arm produces an effect of 0.6 percentage points on this outcome which is insignificant at conventional levels. Both Kai effects remain significant at the 10% level when using Benjamini-Hochberg adjusted p-values. Finally, we find no effects on *Listed at Least Once* for either arm. This outcome captures the extensive margin; the null effect is consistent with an intervention that operates on the intensive margin, reinforcing the decisions of students who were already considering teaching careers rather than attracting new ones.

Panel B shows null effects for both arms on all choice outcomes, indicating that the intervention is effective only among students with prior interest in education. This pattern is consistent with [Ajzenman et al. \(2026\)](#). Panel C, which combines Panels A and B, shows an effect only for the chatbot arm on the *Ranked First* outcome, significant at the 10% level.

For completeness, we also report effects on enrollment, although the SUTVA violation noted above precludes causal interpretation. The point estimate for the human arm is 1.1 percentage points (significant at the 10% level, insignificant after Benjamini-Hochberg correction); the Kai arm’s point estimate is 0.6 percentage points and statistically insignificant. We do not attach causal weight to either estimate.

Overall, our results indicate that Kai produced statistically significant effects on students’ choice of education majors at conventional levels, while the human arm produced smaller, statistically insignificant effects on the same outcomes. The significant effect on *Ranked First* is particularly informative: it shows that Kai induced students to place education programs at the top of their choice set, rather than merely to include them in low positions unlikely to lead to admission under the centralized allocation mechanism. Our findings confirm the pattern documented by [Ajzenman et al. \(2026\)](#): when effective, the *Quiero Ser Profe* intervention

operates only on students with some baseline interest in pursuing an education major. The null effects on the extensive margin further support this interpretation.

The smaller sample size of the human arm raises the question of whether its null result reflects limited statistical precision rather than a true zero effect. The human arm has 4,670 treated and 7,781 control students among education-interested students, a sample size large enough to identify small effects at conventional levels of significance and statistical power, making this possibility less likely. In any case, a conservative reading of our results is that Kai performed at least as well as the human arm, a finding that holds under either interpretation of the human arm null.

7 Mechanisms

We next examine which attributes differ between human counselors and the chatbot, and to what extent they relate to the outcomes of the intervention. Because we cannot link individual conversations to student-level outcomes, we cannot identify which attributes drive effectiveness; instead, we draw on recent methodological advances in the analysis of conversational transcripts (Jabarian and Henkel, 2026) to study several features of student-counselor interactions in both arms and characterize the main differences in how each engages with students.

We first construct a set of measures that capture key features of interactions between students and their assigned counselor, based on anonymized WhatsApp interaction data from both treatment arms. These measures are designed to characterize multiple dimensions of the interactions, including engagement and participation, conversational content and topic coverage, semantic alignment between messages, and the timing and efficiency of exchanges.

To construct these measures, we rely on the underlying message-level data. Each observation in the raw logs is a single time-stamped message, sent either by the student or by the assigned counselor (human or chatbot). This keeps every detail of the interaction, but it cannot be analyzed as is, because chat messages are fragmented (a student may split one message into several parts sent seconds apart).

We therefore preprocess the data through a set of steps that jointly define the analytical sample and the unit of observation in order to make results easier to interpret. First, we restrict the sample to students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction, excluding students who selected the “opt out” option.² These restrictions retain 36,860 messages in the human arm and 37,562 messages in the Kai arm.

Second, we deal with fragmentation. Students often spread one idea over several consecutive messages, which would inflate message counts and add noise unrelated to the quality of the

²This option allowed students to indicate that they did not wish to receive further counseling or be contacted again.

interaction. We therefore group consecutive messages from the same participant sent within 30 seconds of each other, as long as the other participant does not reply in between. We compute all measures (e.g., messages per student) on these grouped units, each of which corresponds to one coherent contribution instead of an arbitrary piece of one.

We construct indicators capturing several conversational attributes, drawing on recent work analyzing conversation dynamics in other settings (Jabarian and Henkel, 2026) as well as focus groups and discussions with the program’s implementers. A detailed description of each outcome is given in Appendix A. Below, we present a brief description.

- *Conversation length*: Total number of messages exchanged in a full conversation. Higher values may reflect greater engagement but could also indicate inefficiencies in information delivery.
- *Topic breadth*: Number of distinct topics discussed in a full conversation (see description of topic identification below).
- *Topic composition*: The specific topics discussed in a full conversation. Topics were classified into three broad categories and 18 subcategories, aligned with Elige Educar’s program guidelines. The first category, admissions and application process, covers the application process, entry requirements, scholarships and financial aid, university inquiries, administrative inquiries, PAES concerns, and the teaching vocation scholarship *Beca Vocación de Profesor*. The second, career and professional outlook, includes the teaching degree, labor market prospects, curricula, professional internships, accreditation, and program quality, compatibility with work, academic demands, and other careers. The third, vocational guidance, encompasses personal experiences, vocational orientation, and conversational interactions.
- *Factual content share*: Proportion of conversational exchanges containing factual content. We also compute the *subjective content share*, defined analogously for exchanges containing subjective content.
- *Counselor coherence*: Cosine similarity between the semantic embeddings of a subject’s message and the counselor’s subsequent response. Higher values indicate that the counselor’s replies are more semantically aligned with the subject’s input.
- *Subject coherence*: Cosine similarity between the semantic embeddings of a counselor’s question and the student’s subsequent response.
- *Counselor response time*: Time between the end of a student turn and the start of the subsequent counselor turn, computed over consecutive turn pairs, measured in minutes.

- *Conversation duration*: For each student-day, minutes between the start of the first student turn and the end of the last student turn. Daily durations are summed across all observed days to obtain cumulative conversation duration per student.
- *Student-to-student gap*: Time between the end of a student’s previous turn and the start of their next turn within the same day, measured in minutes.
- *Time to the first informative exchange*: Number of minutes elapsed before the first important content appears in a conversation. We classify important content as factual or subjective questions.
- *Empathy*: Composite measure that captures the degree of empathy displayed in the conversation, operationalized through emotional reactions, interpretations, and exploratory responses to the student’s experiences (Chen et al., 2024; Sharma et al., 2020).
- *Motivation*: Composite measure that shows the extent to which the counselor encourages reflection and supports student decision-making through personalized encouragement and targeted questions, thus fostering active engagement in the interaction (Sharma et al., 2020).
- *Informativeness*: Composite measure that captures the extent to which the counselor provides clear, relevant, and actionable information for the student’s decision-making process, including both the delivery of pertinent content and the provision of useful external sources tailored to the student’s specific situation.

Using these metrics, we proceed in four steps, examining in turn whether students engaged at all, how long and how fast the exchanges were, what was discussed, and how well the interaction worked.

Length of the conversation: sustained engagement or inefficiency? We first examine the *length* of conversations across arms. As Figure 2 shows, conditional on a conversation starting, human counselors exchange more messages with students than the chatbot does on average.³ This pattern admits two competing interpretations. Students may have been more engaged with human counselors, which would make those exchanges longer. Alternatively, human counselors may deliver information less efficiently. In that case, shorter chatbot conversations would mean that the chatbot answers the student’s query in fewer turns, and not that the student loses interest.

Breadth of the conversation. To distinguish between these interpretations, we start by comparing the *breadth* of conversations, measured as the number of distinct topics covered.

³This difference is larger using raw messages. Students interacting with human counselors split queries across multiple messages, while near-instantaneous chatbot replies induce more consolidated messages. Aggregating messages into turns mitigates this mechanical difference.

Figure 3 shows the distribution of the number of distinct subtopics per student. The mean is 2.24 for Kai and 1.95 for human counselors, with a median of 2 in both arms. Despite exchanging fewer messages per user (7 with the chatbot versus 8 with human counselors), the chatbot spans the same 18 subtopics as its human counterparts. Moreover, the chatbot matches or exceeds the coverage share of human counselors in 11 of the 18 subcategories, indicating that its advantage is not only in the range of topics addressed but also in the relative intensity with which most of them are discussed. A detailed description of the topic taxonomy construction and the automated classification procedure is provided in Appendix A.

Although thematic breadth is comparable, *topic composition* differs substantially across arms. As Figure 4 shows, conversations with the chatbot concentrate disproportionately on admission-related and logistical topics such as admission requirements, scholarships, and the application process (appearing in 19.0%, 24.6%, and 17.9% of Kai conversations, respectively, compared with 11.2%, 18.0%, and 10.4% of human-counselor conversations), whereas exchanges with human counselors are more heavily weighted toward relational and reflective content, including conversational interactions and the sharing of personal experiences (21.6% and 18.3% of human-counselor conversations, compared with 15.1% and 6.5% of Kai conversations). We complement this evidence with two companion measures, *Factual content share* and *Subjective content share*, which capture the proportion of each conversation devoted to factual as opposed to subjective content. These indices are constructed from a Chilean-Spanish BERT model fine-tuned on a manually labeled sample of approximately 900 counselor messages (see Appendix A). Consistent with the topic-composition patterns above, Figure 5 shows that factual content share is substantially larger in the AI arm (25% versus 15%), while the reverse holds for subjective exchanges (4% versus 11%).

Conversational Dynamism. Conversations also differ in their pace, flexibility, and dynamism. Sustained engagement likely depends on how effective counselors are at capturing and sustaining students’ attention, and chatbots have several structural advantages along these dimensions. To begin with, chatbots are available 24/7. Although human counselors do their best to be present when needed, they work in shifts and the student-to-counselor ratio is substantially higher (45 human counselors versus virtually unlimited chatbot capacity). Even if human counselors were constantly present, they would be mechanically slower to respond: they have to think, type, and consult information to give a proper answer. By contrast, chatbots can answer immediately. We explore this using several metrics.

To start, Figure 6 shows that median conversation duration per student was 1.12 minutes with Kai and 28.23 minutes with human counselors; the corresponding means are 35.06 and 120.26 minutes, respectively. As was the case with the conversation-length indicator, longer conversations admit multiple interpretations. Conversations could be longer because students are more engaged and are receiving more information; alternatively, they could be longer because they are less dynamic. The fact that the number of topics covered by each type of counselor is

similar suggests that the former is unlikely to be the most relevant explanation. We complement this evidence with several additional measures. First, as documented in Figure 7, response times are significantly faster for Kai than for human counselors. Every time a student asks a question, Kai provides an answer immediately, while the median human counselor response time is approximately one minute. Delayed answers are likely to affect student engagement: if an answer takes too long, students may lose focus, increase their own response times, and make the conversation less dynamic (Hwang et al., 2019); even in offline interactions, fast responses signal social connection in conversations (Templeton et al., 2022). We capture this through the *Student-to-student gap*. As Figure 8 shows, while the median gap between two sequential student messages is almost three minutes in conversations with human counselors, it is around one minute in conversations with Kai. Finally, longer conversations could mean that a student has to wait longer to obtain valuable information. We capture this idea by measuring the *Time to the first informative exchange*. Figure 9 shows that, while a median conversation with a human counselor reaches substantive content at approximately three-minutes, the equivalent time in Kai conversations is about 44 seconds. Both arms cover a similar range of topics, so we read these timing gaps as differences in the *dynamism* of the exchange. Kai covers about as much ground as a human counselor, in less time.

Kai’s advantages in dynamism are not necessarily indicators of higher quality. Part of the human delay may reflect counselors’ effort to think through and find the best answer for each query. A natural question is thus whether the quality of the conversation differs across groups, regardless of the specific topics discussed or the speed at which they unfold. We use several measures that capture different aspects of conversational quality, grouped into two categories: informational quality and interpersonal quality.

Chatbot-led conversations exhibit higher coherence than human-led ones in both directions: Kai provides more semantically relevant responses to students’ questions, and students in turn respond more coherently to Kai’s prompts than to those of human counselors (Figure 10). This pattern is consistent with the structural design of LLMs, which are trained to generate responses semantically aligned with prompts; we therefore interpret coherence as evidence that Kai’s information delivery is fluent rather than as a marker of higher counseling quality per se.

Quality of the conversation: interpersonal. We turn to the interpersonal dimension of quality, focusing on *empathy* and *motivational support*, two attributes where humans are expected to hold a comparative advantage, and *informativeness*. As detailed in Appendix A, we recruited trained human raters (Chilean psychology students) who scored a random sample of 900 conversations on these and other dimensions, each on a 1–3 scale, with an additional “not applicable” category coded as 0 for cases in which the conversation did not provide a meaningful opportunity for the counselor to display the corresponding skill. We then trained a hybrid classification model to scale these ratings to the full sample of conversations.

Figure 11 reports the mean quality coefficient in each dimension, on a 1–3 scale. Human

counselors substantially outperform Kai on two key socioemotional dimensions: on motivation, the average score is 1.94 for humans against 1.58 for Kai, and on empathy, the gap is even wider, at 1.62 for humans versus 1.13 for Kai. On the informativeness scale (the extent to which counselors provide clear, relevant and actionable information) Kai performs slightly better than humans, although the difference is small. This pattern is consistent with the message-level content analysis reported above: Kai’s replies are more factual and precise, while human counselors rely more heavily on socioemotional content.

Overall, the evidence reveals a coherent trade-off between informational efficiency and socioemotional depth. Kai does better on the delivery of information and the flow of the conversation. It responds faster, students take less time between turns, substantive content comes up earlier, the same range of topics is covered in fewer messages, and semantic coherence is higher in both directions. Human counselors do better on the interpersonal side. Their conversations are longer and more relational, place more weight on subjective content, and score substantially higher on empathy and motivational support. The two arms occupy different points of this trade-off. We interpret the headline result – that the AI arm produces treatment effects at least as large as those of the human arm – as evidence that, in this setting, the informational-efficiency advantage outweighs the interpersonal shortfall. Interpersonal qualities may still matter. For the outcome we measure (ranking an education major), though, the delivery of information appears to be the more binding margin.

7.1 Human Counselor Heterogeneity

A key potential drawback of interventions that depend heavily on implementer quality stems from variation across human implementers (Al-Ubaydli et al., 2017b, 2020). AI-powered delivery can address this constraint by delivering what Jabarian and Henkel (2026) term controlled variance: homogenizing quality across interactions and eliminating the dispersion observed across human implementers. A first way to explore this is to compare each human counselor’s effectiveness with Kai’s in shifting applications toward education majors. Figure 1 plots, for each counselor, the share of their students who listed an education major as their top choice. The upper tail of the distribution seems to perform remarkably well: more than a third of the counselors reach shares comparable to (or above) Kai’s. On average, however, Kai seems to hold only a modest edge over its human counterparts (not large enough to be significantly different at conventional levels). Its value lies less in raising the mean than in compressing the variance: every student assigned to Kai faces the same quality of interaction, whereas quality among human counselors varies widely.

To further explore this pattern, we then ask how Kai compares not only to the average human counselor but to the full distribution of the 45 counselors in our sample, across each of the conversational attributes defined above. For each attribute, we compute the mean for each

human counselor and for Kai, and rank counselors accordingly. Figure 12 reports the results: each bar represents the average score for one human counselor on the corresponding attribute. Figure 13 provides a summary view across all attributes, reporting how many of the 45 human counselors are statistically above, indistinguishable from, or below Kai.

Starting with *topic breadth*, most human counselors are statistically indistinguishable from Kai, while 13 counselors (29%) cover significantly fewer topics per conversation. Results on content composition confirm that conversations with Kai and with human counselors clearly focus on different types of questions (more factual and more subjective, respectively). Roughly 45% of human counselors cover as much *factual content* as Kai, and 27% devote a similar share of turns to *subjective content*. These results document substantial heterogeneity across human counselors in topic emphasis. The picture becomes sharper on dynamism. Kai responds significantly faster than all 45 counselors, and shows significantly shorter student-to-student gaps and shorter conversations than 80% and 73% of them, respectively. For *time to the first informative exchange*, 25 counselors (56%) take significantly longer than Kai, while the remaining 20 (44%) are statistically indistinguishable. No counselor is significantly faster than Kai, or shows significantly shorter gaps or conversations, on any of these dimensions – a mechanical limitation of human delivery rather than a counseling-quality finding.

Turning to conversational quality, heterogeneity is again the central pattern. On *counselor coherence*, Kai significantly outperforms 93% of human counselors, with the remaining 7% statistically indistinguishable; on *subject coherence*, however, only 44% of counselors fall significantly below Kai, with the remaining 56% statistically indistinguishable from the chatbot. On this latter dimension, the upper half of the human distribution is statistically indistinguishable from Kai. The picture inverts on the interpersonal margin, where human counselors generally outperform Kai, but again not uniformly: 7% of counselors are statistically indistinguishable from Kai on *motivation* and 2% are indistinguishable on *empathy*. On rater-scored *informativeness*, only 11% of human counselors significantly outperform Kai, while the rest are statistically indistinguishable or score significantly lower. Across these quality dimensions, the lower tail of human counselors does not significantly outperform Kai.

These patterns speak to the main point of the exercise. On margins where Kai outperforms the average human counselor (*topic breadth*, *subject coherence*, and *informativeness*), a nontrivial share of counselors, those at the top of the distribution, perform as well as the chatbot, and in some cases better. Likewise, on dimensions where human counselors hold a clear advantage (such as *motivation* and *empathy*), a fraction of counselors at the bottom of the distribution do not outperform Kai. The implementer-quality argument predicts this kind of dispersion. The best counselors match or exceed Kai on most dimensions, but the larger the program, the harder it is to keep the whole workforce at that level. Kai’s quality does not depend on scale. This is why a chatbot that scores below the average human counselor on interpersonal dimensions can still have treatment effects at least as large as those of the human arm.

8 Conclusion

We study whether an AI-powered chatbot can replace trained human counselors in a conversation-based behavioral intervention designed to motivate Chilean high school seniors to consider teaching careers. We do this in a nationwide field experiment that randomized roughly 43,000 students into three arms: an AI chatbot calibrated to mimic the human protocol (Kai), the standard human-counseling arm, and a passive control. Among students with baseline interest in teaching (the target of the intervention), treatment by Kai increased the probability of ranking an education major first by 1.4 percentage points and the share of education programs in the choice set by 1.1 percentage points, both significant at conventional levels. The human arm produced smaller and statistically insignificant point estimates. In our setting, AI-led delivery was at least as effective as our trained human counselors and considerably less expensive per treated student.

The cost gap between the two arms is large. Taking each arm’s point estimate at face value (acknowledging that the human-arm estimate is statistically insignificant), reaching one student costs about \$6.20 in the human arm against \$1.95 in the Kai arm, which translates into roughly \$223 versus \$123 per additional student induced to rank an education major first. Without the one-time development cost (amortized over five years), Kai costs about \$0.95 per contact and \$60 per induced student, roughly four times as cost-effective as counselor-based delivery. We consider these figures conservative, because Kai has no binding capacity constraint and the cost of developing and maintaining such systems will probably keep falling as the technology improves. Taken together with the treatment effects, the cost figures imply that at this scale, and on this margin, the chatbot dominates human delivery.

The mechanisms analysis shows that Kai and human counselors have advantages and disadvantages in several conversational attributes. Conversations with Kai are markedly faster and more dynamic (in terms of median conversation duration, counselor response times, and the average time to the first informative content), are more semantically coherent in both directions (from student to counselor and vice versa), and are typically concentrated on factual content such as admission requirements, scholarships, and program-level information. Human counselors, in turn, dominate on the interpersonal margin: their conversations are longer, lean substantially more on subjective and relational content, and score considerably higher on rater-coded *empathy* and *motivation*. Both arms show similar topic breadth, suggesting that Kai is more efficient in delivering a similar amount of information. The two arms therefore show a trade-off between informational efficiency and socioemotional depth. Since the chatbot’s effects are at least as large as those of the human counselors, fast and coherent delivery of information seems able to make up for the lack of warmth, at least in this setting and for this decision.

Quality also varies widely across the 45 human counselors, on almost every conversational attribute we measure. This matters for the design of interventions meant to scale. With

human counselors, the program delivers an average over a heterogeneous pool, and keeping quality constant at scale would almost certainly raise the cost per counselor to the point where the intervention is no longer practical. Kai has no such variation, since every student receives the same content with the same dynamism. The same variation also means that, on almost every metric, including those where Kai is far ahead of the average counselor, some counselors match or beat Kai. If the typical counselor were as good as the best ones in our sample, the average human conversation would probably be at least as good as Kai’s. However, given how much counselors differ, this is not the relevant comparison for policy. When good conversational implementers are scarce, the controlled variance of AI is an advantage in its own right, separate from whether the average AI conversation beats the average human one.

Our findings are relevant to the debate, prompted in part by [Acemoglu et al. \(2026\)](#), on whether AI should substitute for human workers or complement them. Some recent field evidence focuses on augmentation, where humans keep the core of the task and AI raises their productivity at the margin ([Brynjolfsson et al. 2025](#); [Noy and Zhang 2023](#); [Dell’Acqua et al. 2026](#); [Cui et al. 2025](#)). Our experiment is on the substitution side. A fully autonomous AI agent replaced human implementers in a nationwide intervention, with no human supervision, and on a conservative reading produced similar effects at a lower cost. We read this as direct experimental evidence that full automation is viable for conversation-based behavioral interventions.

This conclusion does not rule out complementarity in related settings or for similar interventions, however. Indeed, our mechanisms analysis shows that trained human counselors keep a comparative advantage in *empathy*, *motivation*, and sustaining subjective and relational exchanges, while AI-led delivery is better on dynamism, breadth of factual coverage, and semantic coherence. A natural next step is a hybrid design, with AI handling the dimensions where it is stronger and human counselors focusing on those where they keep a clear advantage. The question would then become how to combine the comparative advantages of AI and human counselors.

9 Tables

Table 1: Baseline Covariates Summary Statistics

	<i>N</i>	Mean	SD	Min	Max
<i>Panel A. Covariates</i>					
Female	41,432	0.650	0.477	0	1
<i>High school funding type:</i>					
Public	40,801	0.359	0.480	0	1
Subsidized private	40,801	0.572	0.495	0	1
Private	40,801	0.069	0.253	0	1
Technical-vocational school	41,432	0.259	0.438	0	1
Prior interest in education	41,432	0.549	0.498	0	1
GPA score (NEM)	41,432	682.7	161.8	0	1,000
Ranking score	41,432	696.1	172.7	0	1,000
Average math and verbal test	41,432	519.5	219.4	0	987.5
Age	41,432	20.34	4.748	17	76
Chilean nationality	41,432	0.825	0.380	0	1
Indigenous ethnicity	41,432	0.148	0.356	0	1
Metropolitan Region	41,432	0.363	0.481	0	1
Rural municipality	41,432	1.400	0.708	0	3
Family income decile	41,432	4.282	2.542	0	10
<i>Panel B. Outcomes</i>					
Ranked first	41,432	0.155	0.362	0	1
Proportion listed	41,432	0.147	0.305	0	1
At least one listed	41,432	0.258	0.438	0	1
Enrolled	41,432	0.136	0.343	0	1

Note. Summary statistics for the estimation sample of 41,432 students, drawn from 2024 PAES registrants who declared an interest in education or social sciences careers and consented to be contacted by Elige Educar. See Section 4.1 for outcome and covariate definitions.

Table 2: Covariates Balance Check

	Human Mean	Chatbot Mean	Control Mean	Human vs. Control	Chatbot vs. Control	Chatbot vs. Human
Female	0.650 (0.005)	0.651 (0.003)	0.650 (0.004)	0.000 (0.007)	0.001 (0.005)	0.001 (0.006)
Public high school	0.362 (0.005)	0.360 (0.004)	0.357 (0.004)	0.005 (0.007)	0.003 (0.005)	−0.002 (0.006)
Subsidized private high school	0.572 (0.005)	0.572 (0.004)	0.572 (0.004)	0.000 (0.007)	0.000 (0.006)	0.000 (0.007)
Private high school	0.066 (0.003)	0.068 (0.002)	0.071 (0.002)	−0.005 (0.003)	−0.003 (0.003)	0.002 (0.003)
Technical-Vocational school	0.264 (0.005)	0.258 (0.003)	0.256 (0.004)	0.008 (0.008)	0.002 (0.001)	−0.006 (0.007)
GPA score (NEM)	679.2 (1.813)	683.2 (1.226)	684.2 (1.433)	−4.918** (2.248)	−0.984 (1.791)	3.934* (2.146)
Ranking score	692.6 (1.926)	696.4 (1.300)	697.6 (1.521)	−4.997** (2.399)	−1.209 (1.911)	3.788* (2.291)
Average math and verbal test	517.7 (2.355)	518.5 (1.602)	522.0 (1.854)	−4.319 (3.013)	−3.543 (2.434)	0.776 (2.885)
Age	20.40 (0.056)	20.33 (0.036)	20.31 (0.041)	0.090 (0.067)	0.027 (0.052)	−0.064 (0.065)
Chilean nationality	0.823 (0.004)	0.823 (0.003)	0.830 (0.003)	−0.007 (0.005)	−0.007* (0.004)	−0.001 (0.005)
Indigenous ethnicity	0.144 (0.004)	0.148 (0.002)	0.151 (0.003)	−0.007 (0.005)	−0.002 (0.004)	0.004 (0.005)
Metropolitan Region	0.357 (0.005)	0.363 (0.003)	0.366 (0.004)	−0.008 (0.007)	−0.003 (0.005)	0.006 (0.006)
Rural municipality	1.394 (0.007)	1.403 (0.005)	1.400 (0.006)	−0.006 (0.010)	0.003 (0.008)	0.009 (0.009)
Family income decile	4.237 (0.027)	4.267 (0.018)	4.327 (0.021)	−0.090** (0.035)	−0.060** (0.028)	0.029 (0.033)
RI p -value				0.102	0.650	0.469
Observations	8,283	18,727	14,422			

Note. Balance across the three experimental arms for the estimation sample of 41,432 students. Each difference reports the OLS coefficient from a regression of the covariate on a treatment indicator. Columns (4) and (5) use the control group as the omitted category; column (6) uses the human arm as the omitted category. The bottom rows report p -values computed via randomization inference following [Kerwin et al. \(2024\)](#). Robust standard errors are shown in parentheses. See Section 4.1 for variable definitions. * $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.

Table 3: Covariates Balance Check by Interest Group – Panel A: Students interested in education

	Human Mean	Chatbot Mean	Control Mean	Human vs. Control	Chatbot vs. Control	Chatbot vs. Human
Female	0.662 (0.007)	0.655 (0.005)	0.648 (0.005)	0.013 (0.009)	0.006 (0.007)	−0.007 (0.008)
Public high school	0.392 (0.007)	0.392 (0.005)	0.393 (0.006)	−0.001 (0.009)	−0.001 (0.007)	0.000 (0.009)
Subsidized private high school	0.567 (0.007)	0.564 (0.005)	0.559 (0.006)	0.008 (0.009)	0.006 (0.008)	−0.002 (0.009)
Private high school	0.042 (0.003)	0.044 (0.002)	0.048 (0.002)	−0.007* (0.004)	−0.005 (0.003)	0.002 (0.004)
Technical-Vocational school	0.293 (0.007)	0.287 (0.004)	0.292 (0.005)	0.000 (0.009)	−0.002 (0.007)	−0.002 (0.008)
GPA score (NEM)	670.779 (2.392)	674.515 (1.588)	675.661 (1.840)	−4.882 (3.017)	−1.146 (2.431)	3.736 (2.871)
Ranking score	684.484 (2.563)	687.882 (1.701)	689.365 (1.969)	−4.881 (3.232)	−1.483 (2.602)	3.397 (3.076)
Average math and verbal test	498.733 (3.278)	503.028 (2.206)	505.611 (2.536)	−6.878* (4.144)	−2.583 (3.362)	4.295 (3.951)
Age	20.455 (0.078)	20.310 (0.048)	20.279 (0.054)	0.137 (0.095)	0.029 (0.072)	−0.108 (0.092)
Chilean nationality	0.831 (0.005)	0.826 (0.004)	0.838 (0.004)	−0.007 (0.007)	−0.007 (0.005)	0.000 (0.007)
Indigenous ethnicity	0.144 (0.005)	0.155 (0.004)	0.156 (0.004)	−0.012* (0.007)	0.000 (0.005)	0.012* (0.006)
Metropolitan Region	0.330 (0.007)	0.326 (0.005)	0.326 (0.005)	0.002 (0.009)	0.000 (0.007)	−0.003 (0.008)
Rural municipality	1.430 (0.011)	1.430 (0.007)	1.435 (0.008)	−0.011 (0.014)	−0.001 (0.011)	0.010 (0.013)
Family income decile	4.183 (0.036)	4.231 (0.024)	4.278 (0.028)	−0.095** (0.045)	−0.048 (0.037)	0.048 (0.043)
RI p -value				0.237	0.950	0.362
Observations	4,670	10,288	7,781			

Note. Balance across the three experimental arms by declared career interest, for the estimation sample of 41,432 students after excluding students who did not sit the PAES (361) and those who opted out (1,743). Each difference reports the OLS coefficient from a regression of the covariate on a treatment indicator. Columns (4) and (5) use the control group as the omitted category; column (6) uses the human arm as the omitted category. The bottom row report p -values computed via randomization inference following [Kerwin et al. \(2024\)](#). Robust standard errors are shown in parentheses. See Section 4.1 for variable definitions. * $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.

Table 4: Covariates Balance Check by Interest Group – Panel B: Students interested in social sciences

	Human Mean	Chatbot Mean	Control Mean	Human vs. Control	Chatbot vs. Control	Chatbot vs. Human
Female	0.634 (0.008)	0.647 (0.005)	0.651 (0.006)	−0.017* (0.010)	−0.005 (0.008)	0.013 (0.010)
Public high school	0.323 (0.008)	0.320 (0.005)	0.314 (0.006)	0.009 (0.010)	0.005 (0.008)	−0.003 (0.009)
Subsidized private high school	0.578 (0.008)	0.581 (0.005)	0.587 (0.006)	−0.009 (0.010)	−0.006 (0.008)	0.003 (0.010)
Private high school	0.098 (0.005)	0.099 (0.003)	0.099 (0.004)	−0.000 (0.006)	0.000 (0.005)	0.001 (0.006)
Technical-Vocational school	0.224 (0.007)	0.216 (0.004)	0.212 (0.005)	0.012 (0.009)	0.004 (0.007)	−0.009 (0.008)
GPA score (NEM)	690.186 (2.711)	693.740 (1.737)	694.123 (1.987)	−3.937 (3.361)	−0.383 (2.639)	3.554 (3.220)
Ranking score	703.154 (2.880)	706.820 (1.849)	707.303 (2.113)	−4.149 (3.571)	−0.482 (2.808)	3.667 (3.422)
Average math and verbal test	542.166 (3.454)	537.260 (2.326)	541.197 (2.613)	0.969 (4.330)	−3.937 (3.498)	−4.906 (4.163)
Age	20.160 (0.075)	20.163 (0.050)	20.146 (0.055)	0.014 (0.093)	0.017 (0.074)	0.003 (0.090)
Chilean nationality	0.805 (0.007)	0.804 (0.004)	0.813 (0.005)	−0.008 (0.008)	−0.009 (0.006)	−0.001 (0.008)
Indigenous ethnicity	0.142 (0.006)	0.137 (0.004)	0.143 (0.004)	−0.001 (0.007)	−0.006 (0.006)	−0.006 (0.007)
Metropolitan Region	0.391 (0.008)	0.405 (0.005)	0.409 (0.006)	−0.018* (0.010)	−0.004 (0.008)	0.014 (0.010)
Rural municipality	1.340 (0.011)	1.351 (0.007)	1.345 (0.008)	−0.005 (0.014)	0.006 (0.011)	0.011 (0.013)
Family income decile	4.307 (0.044)	4.310 (0.029)	4.384 (0.033)	−0.077 (0.055)	−0.073* (0.044)	0.003 (0.053)
RI p -value				0.313	0.741	0.552
Observations	3,613	8,439	6,641			

Note. Balance across the three experimental arms by declared career interest, for the estimation sample of 41,432 students after excluding students who did not sit the PAES (361) and those who opted out (1,743). Each difference reports the OLS coefficient from a regression of the covariate on a treatment indicator. Columns (4) and (5) use the control group as the omitted category; column (6) uses the human arm as the omitted category. The bottom row report p -values computed via randomization inference following [Kerwin et al. \(2024\)](#). Robust standard errors are shown in parentheses. See Section 4.1 for variable definitions. * $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.

Table 5: Intent-to-Treat Effects on Application and Enrollment Outcomes

	(1)	(2)	(3)	(4)
	Ranked First	Proportion	Listed at Least Once	Enrollment
<i>Panel A. Students interested in education</i>				
Treat Chatbot	0.014**† (0.006)	0.011**† (0.005)	0.007 (0.006)	0.006 (0.005)
Treat Human	0.011 (0.007)	0.006 (0.006)	0.000 (0.007)	0.011* (0.006)
<i>p</i> -value (Chatbot = Human)	0.670	0.424	0.284	0.478
<i>R</i> ²	0.248	0.297	0.427	0.346
Observations	22,739			
<i>Panel B. Students interested in social sciences (not education)</i>				
Treat Chatbot	0.001 (0.002)	0.001 (0.002)	0.004 (0.004)	0.001 (0.002)
Treat Human	−0.002 (0.003)	−0.001 (0.002)	−0.001 (0.005)	−0.001 (0.003)
<i>p</i> -value (Chatbot = Human)	0.406	0.412	0.357	0.345
<i>R</i> ²	0.018	0.038	0.059	0.027
Observations	18,693			
<i>Panel C. Full estimation sample</i>				
Treat Chatbot	0.006* (0.003)	0.005 (0.003)	0.004 (0.004)	0.002 (0.003)
Treat Human	0.004 (0.004)	0.001 (0.004)	−0.003 (0.005)	0.005 (0.004)
<i>p</i> -value (Chatbot = Human)	0.535	0.341	0.155	0.438
<i>R</i> ²	0.248	0.292	0.359	0.278
Observations	41,432			

Notes: OLS estimates from equation (1). Panels A and B restrict the estimation sample to students who, at PAES registration, declared interest in education or in social sciences (but not education), respectively; Panel C pools both groups. The dependent variables are: (1) an indicator for listing an education program first in the student’s choice set; (2) the share of education programs in the choice set; (3) an indicator for listing at least one education program; and (4) an indicator for enrolling in an education program. All specifications include the full set of individual controls (GPA score, ranking score, average PAES score, age, school type, nationality, and indigenous status) as well as fixed effects for randomization strata, family income decile, Metropolitan Region, rurality, and missing-value indicators. “ p -value (Chatbot = Human)” reports the test of equality between the two treatment coefficients. Robust standard errors are in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. † indicates that the coefficient remains significant at the 10% level after applying the [Benjamini and Hochberg \(1995\)](#) false discovery rate correction within each treatment arm, treating the outcomes in each panel as the family of hypotheses ($m = 4$). See [Benjamini and Hochberg \(1995\)](#) and [Anderson \(2008\)](#).

10 Figures

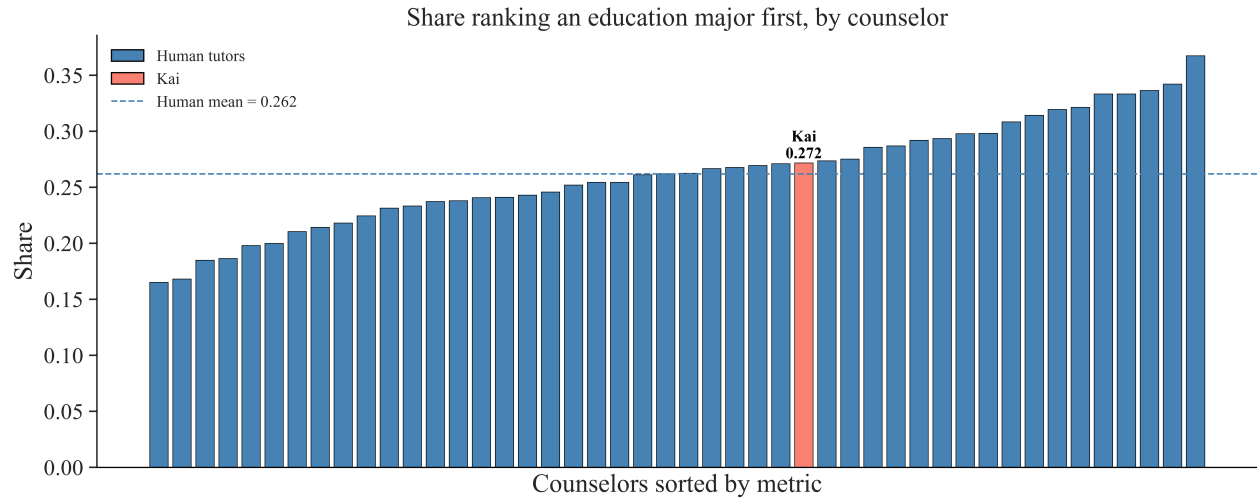


Figure 1: Share of students ranking an education major first, by counselor

Notes: The figure reports the mean of *Ranked first*, an indicator equal to one if the student listed an education major as their first-choice university-program combination, for each human counselor (blue bars) and Kai (salmon bar). We include students who declared an interest in education careers at PAES registration, leaving out those who opted out or did not sit the PAES. The outcome is measured for each student, whether or not we observe a conversation. The dashed horizontal line is the simple average of the human-counselor means shown.

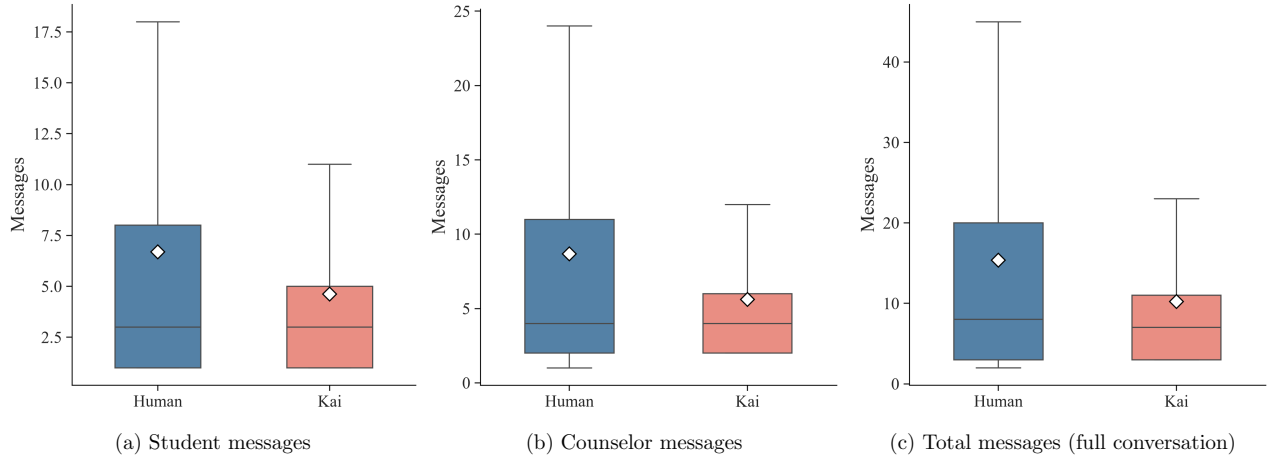


Figure 2: Number of messages per student, by speaker and total

Notes: The figure shows the distribution of the number of messages per student across treatment arms, broken down by speaker. Panel (a) counts messages sent by the student, Panel (b) messages sent by the counselor (chatbot or human), and Panel (c) the total number of messages exchanged by the student and counselor. There is one observation per student, and messages are counted over all days we observe. A message groups consecutive entries sent by the same participant within 30 seconds, with no reply from the other side in between. We include students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction, and exclude those who opted out of further contact. Boxes show the interquartile range (IQR) and the median, and the white diamond is the mean. Whiskers reach the most extreme observation within $1.5 \times \text{IQR}$ of the box. Outliers are not shown. For a more detailed and technical description of the data construction and classification procedures, see Appendix A.

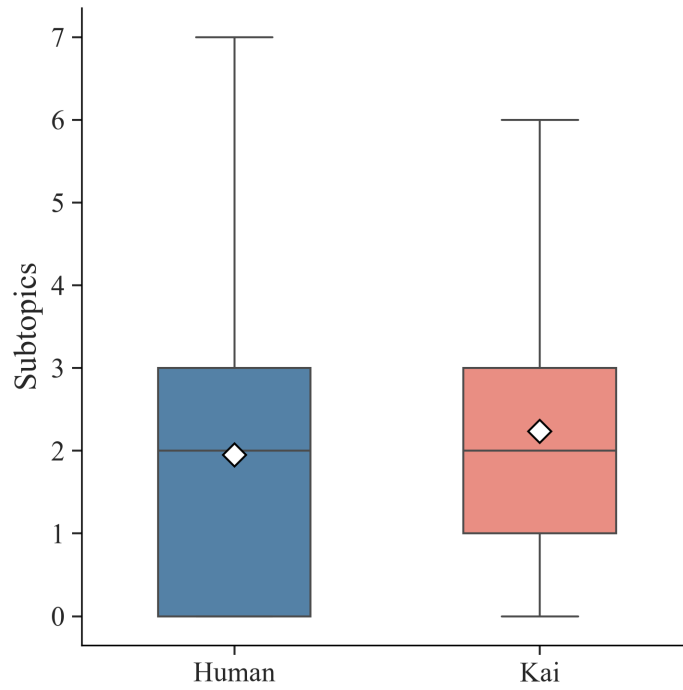


Figure 3: Topic breadth by treatment arm

Notes: The figure shows the distribution of topic breadth across treatment arms, measured as the number of distinct subtopics detected in each student’s conversation, pooling content across all observed days. Each observation is a student. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. Each box spans the interquartile range (IQR) and marks the median, with the mean as a white diamond. Whiskers go out to the most extreme observation within $1.5 \times \text{IQR}$ of the box, and outliers are not plotted. For a more detailed and technical description of the data construction and classification procedures, see Appendix A.

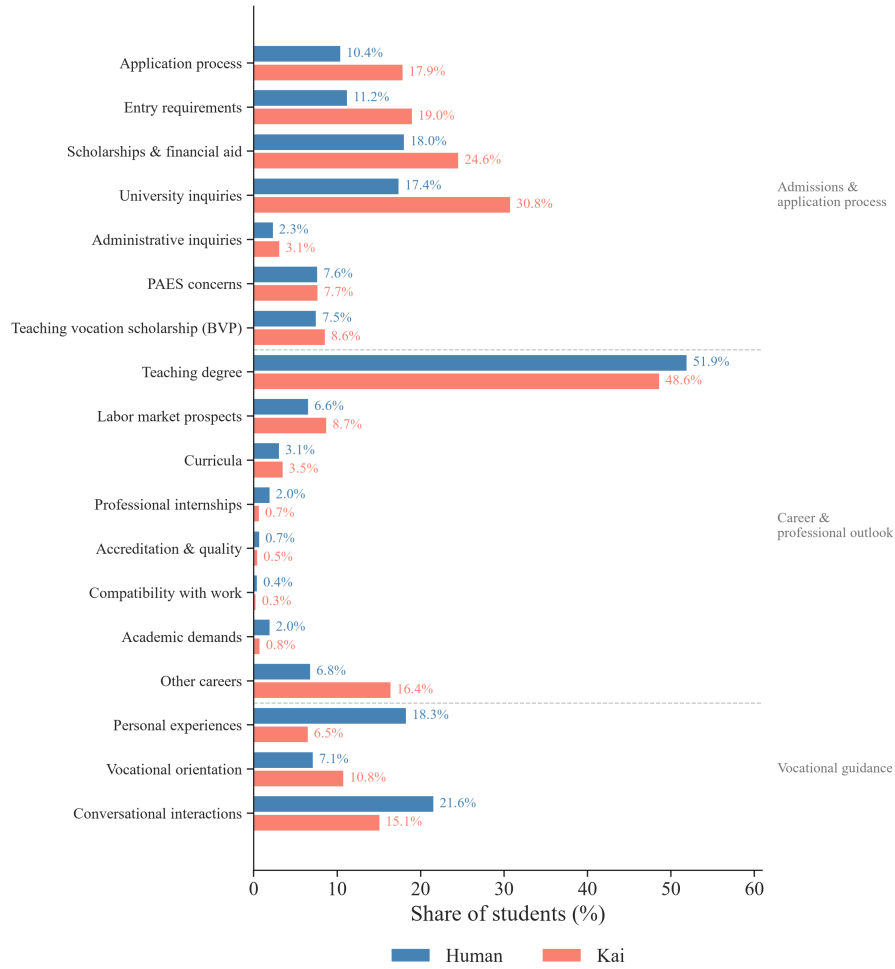


Figure 4: Prevalence of Conversation Topics by Treatment Arm

Notes: The figure reports, for each treatment arm, the share of students who addressed each subcategory. A subcategory is coded 1 for a student if at least one of their messages across all observed days addresses that topic. Topics are grouped into three macro-categories: admissions and application process (subcategories 1–7), career and professional outlook (8–15), and vocational guidance (16–18). The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. For a more detailed and technical description of the data construction and classification procedures, see Appendix A.

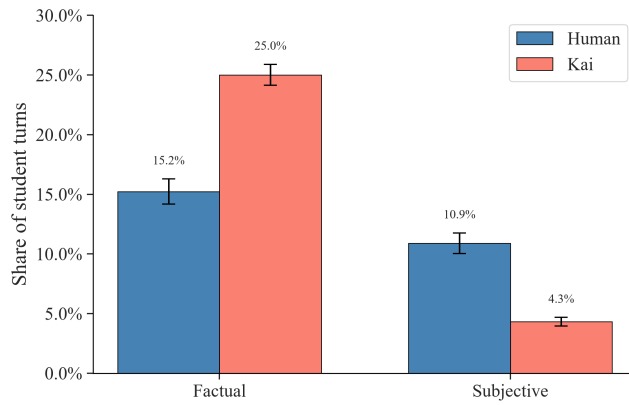


Figure 5: Share of student turns classified as factual or subjective, by treatment arm

Notes: The figure reports, for each treatment arm, the mean share of each student’s turns classified as factual or subjective. Each share is calculated using all observed student turns across all days. Classifications are based on the fine-tuned text classifier applied to student turns. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed student turn; students who opted out of further contact are excluded. Error bars indicate 95% confidence intervals. For a more detailed and technical description of the data construction and classification procedures, see Appendix A.

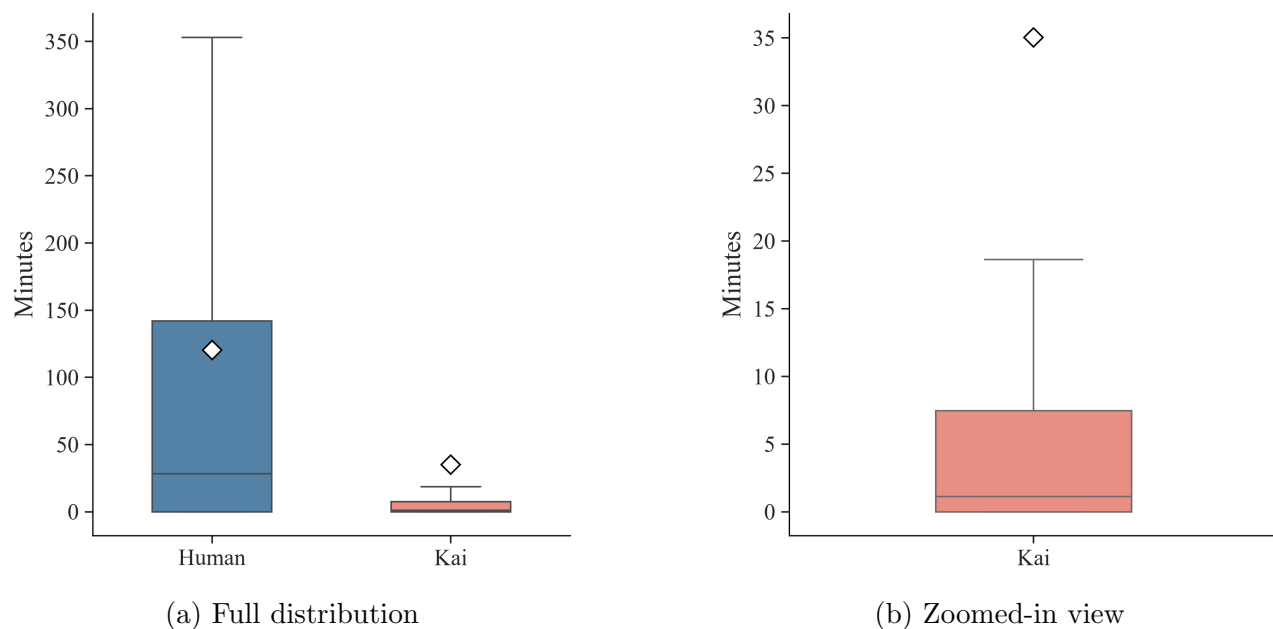


Figure 6: Conversation duration

Notes: The figure shows the distribution of total conversation duration per student across treatment arms, measured in minutes. Duration per student is the sum, over observed days, of the time between the start of the student's first turn and the end of their last turn that day (pauses included, gaps between days excluded). Each observation is a student. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. Boxes show the interquartile range (IQR) and the median; the white diamond marks the mean; whiskers extend to the most extreme observations within $1.5 \times \text{IQR}$ of the box, and outliers are suppressed for readability. For a more detailed and technical description of the data construction and classification procedures, see Appendix A.

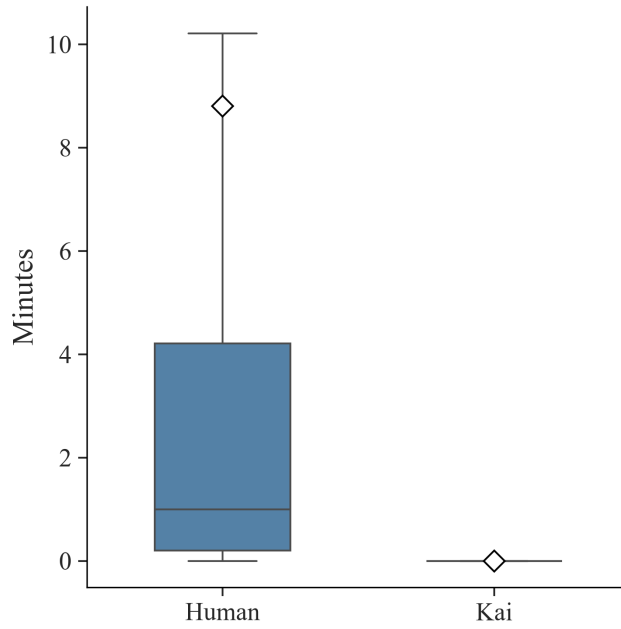


Figure 7: Counselor Response Time by Treatment Arm

Notes: The figure shows the distribution of counselor response times across treatment arms, measured in minutes from the end of a student turn to the start of the immediately following counselor turn within the same student-day. Responses to the first student turn of each day are excluded. Each observation is an eligible student-to-counselor transition. Kai's response times are set to zero. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. Boxes show the interquartile range (IQR) and the median; the white diamond marks the mean; whiskers extend to the most extreme observations within $1.5 \times \text{IQR}$ of the box, and outliers are suppressed for readability. For a more detailed and technical description of the data construction and classification procedures, see Appendix A.

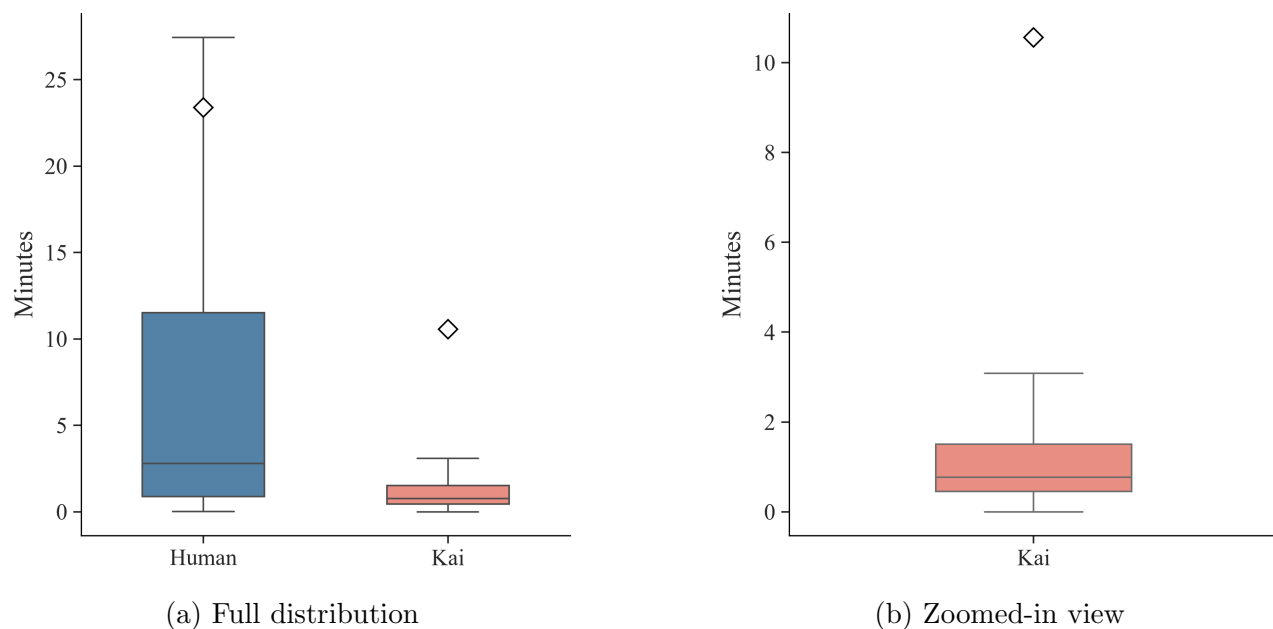


Figure 8: Student-to-student gap

Notes: The figure shows the distribution of student-to-student gaps across treatment arms, measured in minutes from the end of a student's previous turn to the start of their next turn within the same student-day. Each observation is the gap between consecutive student turns. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. Boxes show the interquartile range (IQR) and the median; the white diamond marks the mean; whiskers extend to the most extreme observations within $1.5 \times \text{IQR}$ of the box, and outliers are suppressed for readability. For a more detailed and technical description of the data construction and classification procedures, see Appendix A.

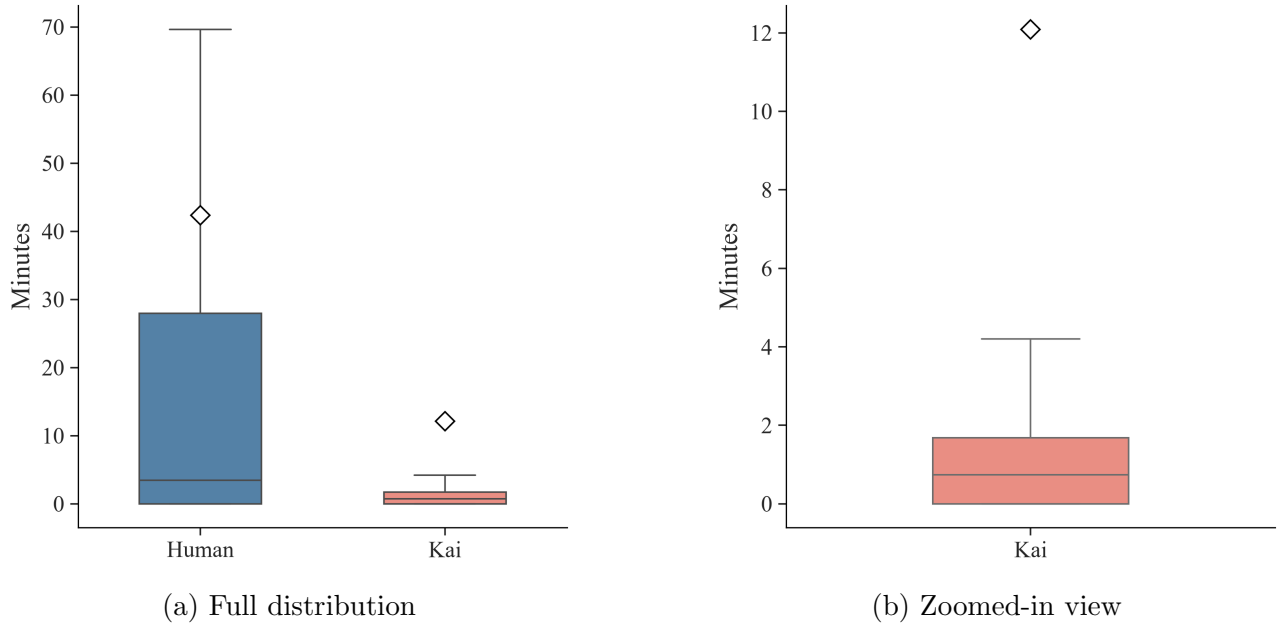


Figure 9: Time to the first informative exchange

Notes: The figure shows the distribution of time until the first informative student turn across treatment arms, measured in minutes from the start of the first student turn to the start of the first student turn classified as factual or subjective within the same student-day. Each observation is a student-day with at least one such informative turn. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. Boxes show the interquartile range (IQR) and the median; the white diamond marks the mean; whiskers extend to the most extreme observations within $1.5 \times \text{IQR}$ of the box, and outliers are suppressed for readability. For a more detailed and technical description of the data construction and classification procedures, see Appendix A.

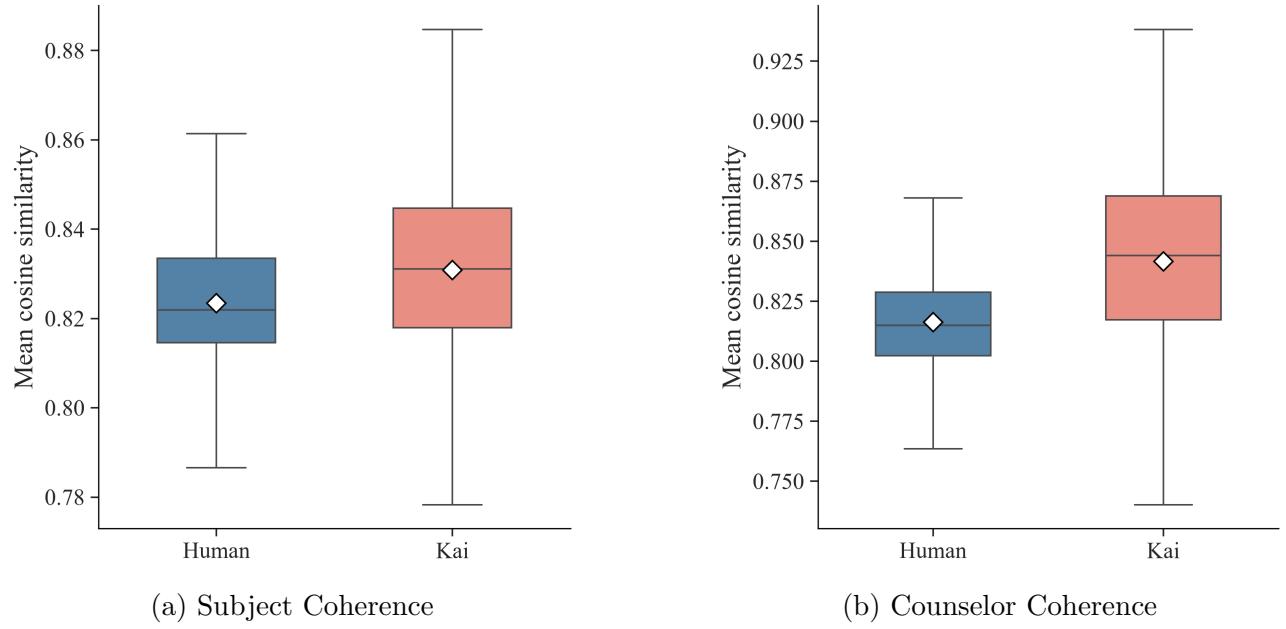


Figure 10: Subject and counselor coherence per student

Notes: The figure shows the distribution of subject coherence (left) and counselor coherence (right), measured using cosine similarity for counselor-to-student and student-to-counselor prompt-and-reply pairs, respectively. Each observation is a student: coherence is calculated as the mean similarity across all eligible pairs involving that student across all observed days. Pairs containing automatic button-click acknowledgements are excluded. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. Boxes show the interquartile range (IQR) and the median; the white diamond marks the mean; whiskers extend to the most extreme observations within $1.5 \times \text{IQR}$ of the box, and outliers are suppressed for readability. For a more detailed and technical description of the data construction and classification procedures, see Appendix A.

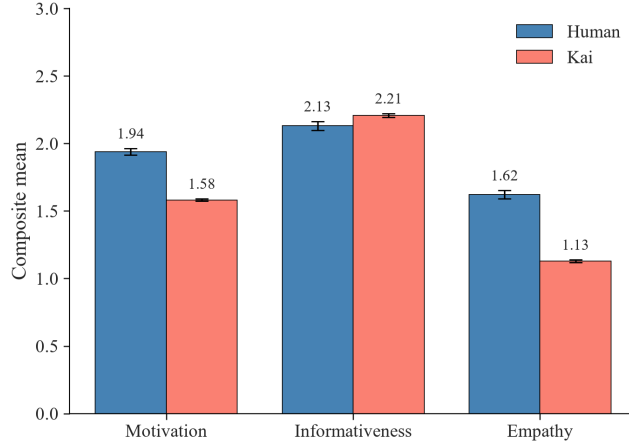


Figure 11: Quality means by treatment arm

Notes: The figure reports, for each treatment arm, the mean predicted quality coefficient for motivation, informativeness, and empathy. For each dimension, the training target is the mean of its applicable sub-dimension scores on a 1–3 scale. These composites are predicted directly using a separate regressor for each dimension, selected between random forest and gradient boosting by 5-fold cross-validated MAE. The selected models are random forest for motivation and empathy, and gradient boosting for informativeness. Models are trained on hand-labeled conversations, and quality predictions are reported only for conversations classified as applicable by an upstream model. Each observation is a student, with conversation content pooled across all observed days. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. Bars indicate the arm-level mean and whiskers the 95% confidence interval. For a more detailed and technical description of the labeling, classification, and aggregation procedures, see Appendix A.

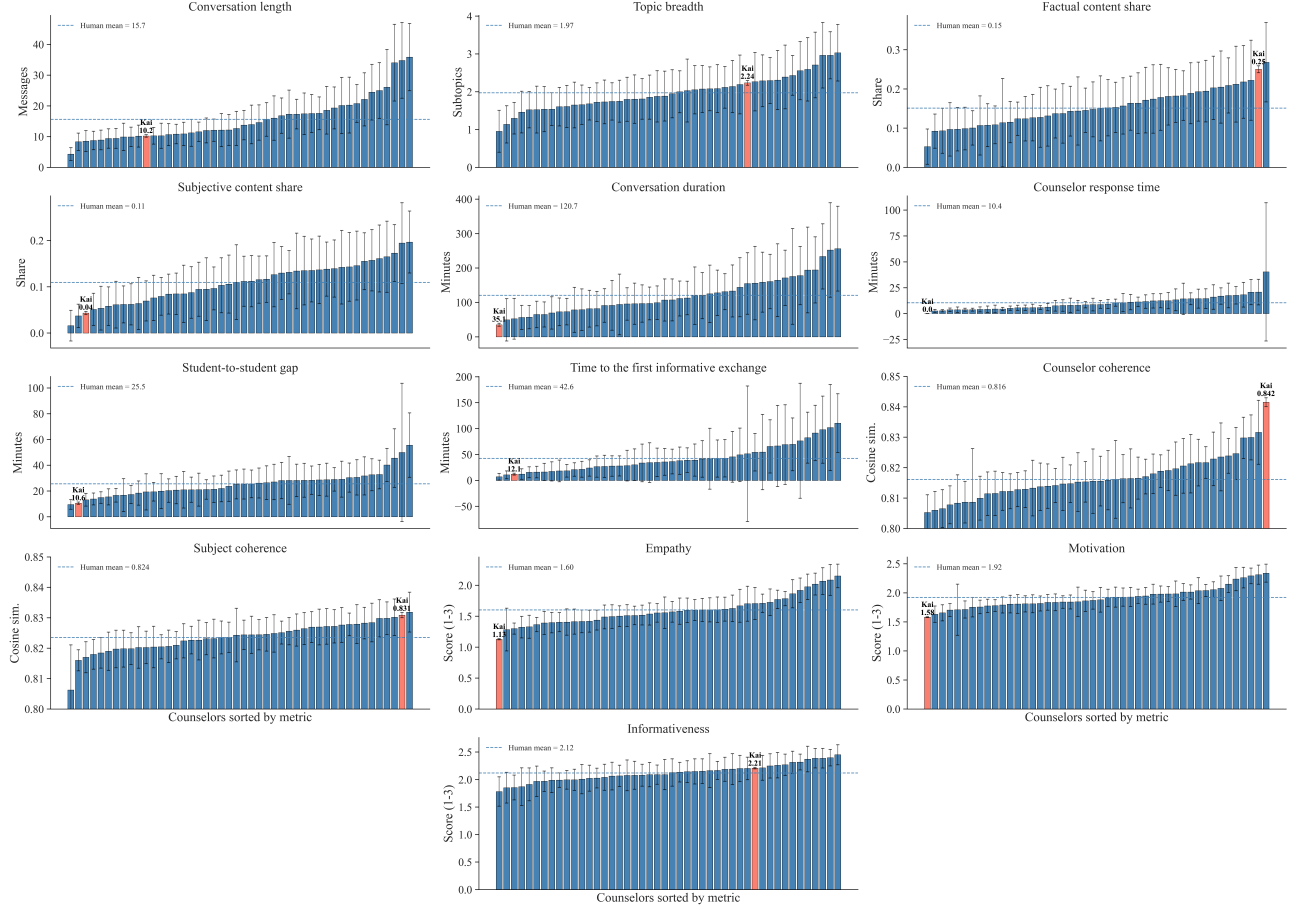


Figure 12: Counselor-level performance distribution and Kai’s position across conversation metrics

Notes: The figure reports the mean of each metric for each human counselor (blue bars, sorted ascending) together with Kai’s mean (salmon bar, inserted in the same sort order). Message counts and factual and subjective shares are calculated per student across all observed days. Counselor response time is the time between the end of a student turn and the start of the counselor turn that follows it on the same day; we drop responses to the student’s first turn of each day. The student-to-student gap is the time from the end of one student turn to the start of the same student’s next turn that day, keeping only strictly positive gaps. All eligible responses and gaps are weighted equally. Conversation duration is measured from the start of the first student turn to the end of the last student turn within each student-day, then summed across all observed days for each student. Time to the first informative exchange is measured from the start of the first student turn to the start of the first factual or subjective student turn within each student-day, excluding days without an informative turn. Counselor and subject coherence are mean cosine similarities across eligible student-to-counselor and counselor-to-student message pairs, respectively, within each student, excluding automatic button-click acknowledgements. Composite quality scores are predicted regressor outputs for students whose pooled conversation content is classified as applicable. Topic breadth is the count of distinct subcategories detected across each student’s observed days. Whiskers show 95% confidence intervals. Human counselors are shown if they have at least two valid observations for the corresponding metric. The dashed horizontal line marks the simple average of the displayed human-counselor means. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. For a more detailed and technical description of the data construction, classification, and aggregation procedures, see Appendix A.

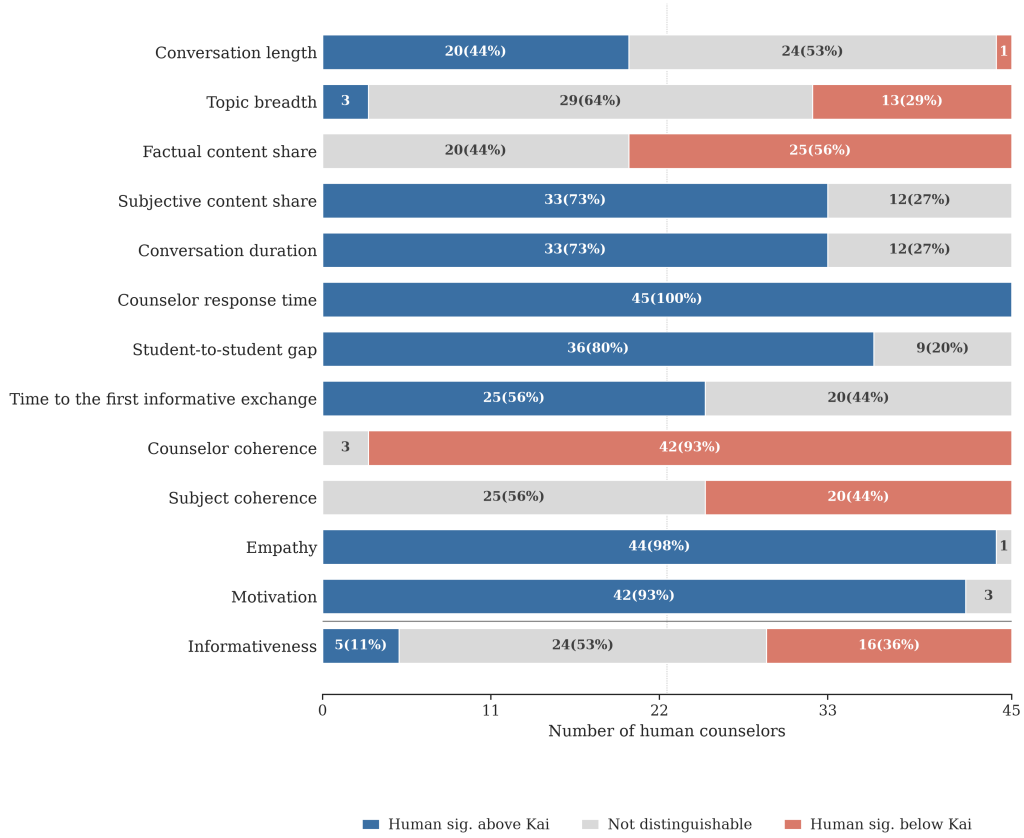


Figure 13: Comparison between Kai and human counselors

Notes: The figure reports, for each metric, the number of human counselors classified as statistically above Kai, indistinguishable from Kai, or below Kai. For each counselor, we compare their observations with Kai’s using a two-sample t-test with equal variances ($\alpha = 5\%$), without adjusting for multiple comparisons. “Above” and “below” refer to the numeric sign of the difference. Metrics are defined as in Figure 12. We compare all 45 human counselors, and a test needs at least two valid observations. The sample includes students who declared an interest in education careers at PAES registration, sat the PAES, and had at least one observed interaction; students who opted out of further contact are excluded. For a more detailed and technical description of the metric construction, classification, and inference procedures, see Appendix A.

References

- Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy*, 40.
- Acemoglu, D., Autor, D., and Johnson, S. (2026). Building pro-worker artificial intelligence. NBER Working Paper 34854, National Bureau of Economic Research.
- Ajzenman, N., Elacqua, G., Hincapié, D., Jaimovich, A., Boo, F. L., Paredes, D., and Román, A. (2021). Career choice motivation using behavioral strategies. *Economics of Education Review*, 84:102173.
- Ajzenman, N., Elacqua, G., Jaimovich, A., and Pérez-Núñez, G. (2026). Humans versus chatbots: Scaling up behavioral interventions to reduce teacher shortages. *Journal of Political Economy Microeconomics*, 4(2).
- Ajzenman, N., Elacqua, G., Marotta, L., and Westh Olsen, A. S. (2023). Order effects and employment decisions: Experimental evidence from a nationwide program. IDB Publications (Working Papers) 11541, Inter-American Development Bank.
- Al-Ubaydli, O., Lee, M. S., List, J. A., Mackevicius, C. L., and Suskind, D. (2021). How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural Public Policy*, 5(1):2–49.
- Al-Ubaydli, O., List, J. A., LoRe, D., and Suskind, D. (2017a). Scaling for economists: Lessons from the non-adherence problem in the medical literature. *Journal of Economic Perspectives*, 31(4):125–144.
- Al-Ubaydli, O., List, J. A., and Suskind, D. (2017b). What can we learn from experiments? Understanding the threats to the scalability of experimental results. *American Economic Review*, 107(5):282–286.
- Al-Ubaydli, O., List, J. A., and Suskind, D. (2020). The science of using science: Towards an understanding of the threats to scaling experiments. *International Economic Review*, 61(4):1387–1409.
- Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495.
- Ash, E. and Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, 15(1):659–688.
- Ávalos, B. and Bellei, C. (2019). Recent education reforms in chile: How much of a departure from market and new public management systems? In *Politics of education in Latin America*, pages 43–71. Brill.

- Bacolod, M. P. (2007). Do alternative opportunities matter? The role of female labor markets in the decline of teacher quality. *Review of Economics and Statistics*, 89(4):737–751.
- Benjamini, Y. and Hochberg, Y. (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.
- Blattman, C., Jamison, J. C., and Sheridan, M. (2017). Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia. *American Economic Review*, 107(4):1165–1206.
- Boyd, D., Grossman, P., Hammerness, K., Lankford, H., Loeb, S., Ronfeldt, M., and Wyckoff, J. (2012). Recruiting effective math teachers: Evidence from New York City. *American Educational Research Journal*, 49(6):1008–1047.
- Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J. (2005). The draw of home: How teachers’ preferences for proximity disadvantage urban schools. *Journal of Policy Analysis and Management*, 24(1):113–132.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942.
- Carrell, S. E. and Sacerdote, B. (2017). Why do college-going interventions work? *American Economic Journal: Applied Economics*, 9(3):265–291.
- Castro-Zarzur, R., Espinoza, R., and Sarzosa, M. (2022). Unintended consequences of free college: Self-selection into the teaching profession. *Economics of Education Review*, 89:102260.
- Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2023). Spanish pre-trained bert model and evaluation data.
- Chen, R., Chen, H., Kulkarni, A., Lin, E., Pang, L., Tadimeti, D., Shin, J., and Hirschberg, J. (2024). Detecting empathy in speech. *Interspeech 2024*.
- Chen, Y., Cheng, K., Li, W.-W., et al. (2025). An LLM chatbot to facilitate primary-to-specialist care transitions: A randomized controlled trial. *Nature Medicine*.
- Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics*, 126(4):1593–1660.
- Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–79.

- Chopra, F., Haaland, I., Röver, N., and Roth, C. (2026). Evaluating behavioral interventions at scale with AI. ECONtribute Discussion Paper 385.
- Christian, A., Ronfeldt, M., and Zafar, B. (2024). College students and career aspirations: Nudging student interest in teaching. Working Paper 32641, National Bureau of Economic Research.
- Clotfelter, C. T., Glennie, E., Ladd, H. F., and Vigdor, J. L. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, 92(5–6):1352–1370.
- Corcoran, S. P., Evans, W. N., and Schwab, R. M. (2004). Changing labor-market opportunities for women and the quality of teachers, 1957–2000. *American Economic Review*, 94(2):230–235.
- Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., and Zamora, P. (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Quarterly Journal of Economics*, 128(2):531–580.
- Cui, Z. K., Demirer, M., Jaffe, S., Musolff, L., Peng, S., and Salz, T. (2025). The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers. *Management Science*.
- Davis, J. M., Guryan, J., Hallberg, K., and Ludwig, J. (2017). The economics of scale-up. Working Paper 23925, National Bureau of Economic Research.
- de Ree, J., Muralidharan, K., Pradhan, M., and Rogers, H. (2018). Double for nothing? Experimental evidence on an unconditional teacher salary increase in Indonesia. *Quarterly Journal of Economics*, 133(2):993–1039.
- Dell’Acqua, F., McFowland, E., Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Kraymer, L., Candelon, F., and Lakhani, K. R. (2026). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science*, 37(2):403–423.
- DellaVigna, S. and Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1):81–116.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.
- Elacqua, G., Hincapié, D., Hincapié, I., and Montalva, V. (2022). Can financial incentives help disadvantaged schools to attract and retain high-performing teachers? Evidence from Chile. *Journal of Policy Analysis and Management*, 41(2):603–631.
- Elacqua, G., Soares, S., Marotta, L., Martínez, M., Montalva, V., Westh Olsen, A. S., and Méndez, C. (2020). El problema de la escasez de docentes en Latinoamérica y las políticas para enfrentarlo. Technical report, Inter-American Development Bank.

- Falch, T. (2011). Teacher mobility responses to wage changes: Evidence from a quasi-natural experiment. *American Economic Review*, 101(3):460–465.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.
- Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15.
- Geiecke, F. and Jaravel, X. (2024). Conversations at scale: Robust AI-led interviews. Forthcoming, *Review of Economic Studies*. SSRN Working Paper 4974382.
- Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.
- Han, S. W., Borgonovi, F., and Guerriero, S. (2018). What motivates high school students to want to be teachers? The role of salary, working conditions, and societal evaluations about occupations in a comparative perspective. *American Educational Research Journal*, 55(1):3–39.
- Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., and Jacobson, N. C. (2025). Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI*, 2(4):AIoa2400802.
- Hoxby, C. M. and Leigh, A. (2004). Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States. *American Economic Review*, 94(2):236–240.
- Hwang, S. Y., Khojasteh, N., and Fussell, S. R. (2019). When delayed in a hurry: Interpretations of response delays in time-sensitive instant messaging. *Proceedings of the ACM on Human-Computer Interaction*, 3(GROUP):1–19.
- Ide, E. and Talamàs, E. (2024). Artificial intelligence in the knowledge economy. Working paper, IESE Business School.
- Jabarian, B. and Henkel, L. (2026). Voice AI in firms: A natural field experiment on automated job interviews. Working paper, University of Chicago Booth School of Business and Erasmus University Rotterdam.
- Jackson, C. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics*, 27(2):213–256.
- Karnani, M. (2024). *Essays on Empirical Matching Systems*. Massachusetts Institute of Technology.
- Kerwin, J. T., Rostom, N., and Sterck, O. (2024). Striking the right balance: Why standard balance tests over-reject the null, and how to fix it. IZA Discussion Paper 17217, Institute of Labor Economics (IZA).

- Korinek, A. and Suh, D. (2024). Scenarios for the transition to AGI. Working Paper 32255, National Bureau of Economic Research.
- Kraft, M. A., Blazar, D., and Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, 88(4):547–588.
- Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. Sage, 4 edition.
- Mehri, S. and Eskenazi, M. (2020). USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707. Association for Computational Linguistics.
- Ministerio de Educación, Centro de Estudios and Elige Educar (2025). Disponibilidad y necesidad de horas docentes en el sistema escolar chileno: Análisis según idoneidad y especialidad año 2024. Technical report, Ministerio de Educación de Chile, Santiago, Chile.
- Muralidharan, K. and Niehaus, P. (2017). Experimentation at scale. *Journal of Economic Perspectives*, 31(4):103–124.
- Naeem, M., Smith, T., and Thomas, L. (2025). Thematic analysis and artificial intelligence: A step-by-step process for using chatgpt in thematic analysis. *International Journal of Qualitative Methods*, 24:1–18.
- Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381:187–192.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.
- Pérez-Núñez, G. (2020). *Building a Teacher Pipeline: Evidence from a Merit-Based Tuition Scholarship*. PhD thesis, Stanford University. Chapter in unpublished dissertation: *Advancing Equity: The Effects of Higher Education Interventions for Improving Underserved Students’ Educational Outcomes*, pp. 112–158.
- Pérez-Rosas, V., Wu, X., Resnicow, K., and Mihalcea, R. (2019). What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

- Perryman, J. and Calvert, G. (2020). What motivates people to teach, and why do they leave? Accountability, performativity and teacher retention. *British Journal of Educational Studies*, 68(1):3–23.
- Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3982–3992.
- Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.
- Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2):247–252.
- Sharma, A., Miner, A. S., Atkins, D. C., and Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.
- Staiger, D. O. and Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, 24(3):97–118.
- Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune BERT for text classification? In *Proceedings of CCL 2019*, pages 194–206.
- Tao, C., Mou, L., Zhao, D., and Yan, R. (2018). RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 722–729.
- Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., and Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, 119(4):e2116915119.
- UNDP, U. N. D. P. (2023). Evaluación del sistema de apoyo formativo y del sistema de reconocimiento y promoción de la ley 20.903. Technical report, United Nations Development Programme (UNDP). Accessed: 2025-03-31.
- Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, 18(6):3045–3089.
- Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual E5 text embeddings: A technical report. Technical report, arXiv:2402.05672.
- Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45.

Appendix

A Technical Appendix

This appendix documents the construction of the conversational measures used in the Mechanisms section of the paper. We work only with the anonymized WhatsApp logs. Some measures are computed per *session* (one student on one date) and others per *conversation*, which pools all of a student’s sessions. The pipeline is implemented in Python, relying on `transformers` (Wolf et al., 2020), `sentence-transformers` (Reimers and Gurevych, 2019), `PyTorch` (Paszke et al., 2019), and `scikit-learn` (Pedregosa et al., 2011).

A.1 From raw messages to analytic turns

A.1.0.1 Message-burst collapsing. A single WhatsApp message is often a fragment: a greeting, then the question, then a clarification, each sent separately. Counting fragments would inflate message counts and distort timing and content measures. We merge them into *turns*. Consecutive messages from the same participant belong to one turn if (i) fewer than 30 seconds separate them and (ii) the other participant has not written in between. The turn’s text is the messages joined by line breaks, and it starts with the first message and ends with the last. Everything downstream (message counts, response times, embeddings) is computed on turns.

A.1.0.2 Units of observation. A *session* is one student on one date, identified by (`student_id`, `date`). A student contacted on two dates has two sessions. Timing measures (duration, response time) are defined per session. A *conversation* is identified by `student_id` alone and pools all of the student’s sessions. We compute content, coherence, and quality indicators per conversation, which is also the level at which human raters scored quality (see Section A.6).

A.1.0.3 Sample restrictions. We drop students who opted out, together with all their conversations. We also keep only sessions (student-days) with at least one student turn and at least one counselor turn, which removes days when the student did not reply.

A.2 Interaction volume and timing

All timing measures are computed on burst-collapsed turns.

- *Conversation length*: total turns exchanged; we also report student and counselor turn counts separately.

- *Conversation duration*: For each student-day, minutes between the start of the first student turn and the end of the last student turn. Daily durations are summed across all observed days to obtain cumulative conversation duration per student.
- *counselor response time*: for every counselor turn whose immediately preceding turn in the same session is a student turn, the minute gap between the student’s turn end and the counselor’s turn start. For the Kai arm, this gap is set to zero: API-level latency is negligible at the minute resolution at which the human arm operates and would otherwise contribute only machine-precision jitter. Since Human tutors have working hours, responses to the first student turn of each day are excluded to avoid compute response to contacts outside working time.
- *Student-to-student gap*: for every non-initial student turn, the minutes elapsed since the end of the previous student turn within a student-day. This captures how long the student remains engaged in the dialogue.

A.3 Evaluation metrics

Throughout Sections A.4 and A.6 we report the same battery of out-of-sample classification metrics, defined below for reference.

- **Accuracy**: share of predictions that exactly match the true label. It is an unweighted average over examples; it is *not* robust to class imbalance and can be misleadingly high when one class dominates.
- **Macro-F1**: an imbalance-robust counterpart to accuracy. For each class c , let TP_c , FP_c and FN_c denote the number of true positives, false positives and false negatives. *Precision* and *recall* for class c are

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c},$$

and the per-class $F1_c$ is their harmonic mean, $F1_c = 2P_cR_c/(P_c + R_c)$; intuitively, $F1_c$ is high only when the classifier both identifies most true class- c examples *and* does not over-predict the class. Macro-F1 averages the per-class F1 scores,

$$\text{macroF1} = \frac{1}{K} \sum_{c=1}^K F1_c,$$

so every class counts the same no matter its size. Accuracy, which here equals micro-F1, weights *examples* instead of classes and can be driven almost entirely by the majority class. Take a dimension on which 90% of conversations are not applicable. Predicting

“not applicable” every time gives 90% accuracy and no information about the other 10%. Under macro-F1 the minority class keeps a weight of $1/K$, so this classifier scores poorly. We select models by macro-F1 throughout the appendix.

- **Regression metrics.** For continuous (composite) targets we report mean absolute error ($\text{MAE} = \frac{1}{n} \sum_i |y_i - \hat{y}_i|$), root-mean-squared error ($\text{RMSE} = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$), the coefficient of determination $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, and Pearson and Spearman correlations (r , ρ) between predictions and labels. MAE and RMSE are in composite units (1 to 3). R^2 compares the model’s squared errors with those of predicting the sample mean, and is negative when the model does worse. Pearson’s r captures linear association. Spearman’s ρ uses ranks, which suits an ordinal scale where the relationship may be monotone but not linear.

A.4 Content classification

A.4.0.1 Supervised factual / subjective. Each student message is labeled *factual*, *subjective*, both, or neither. Factual messages ask for or give verifiable information (scores, deadlines, requirements), and subjective ones express an opinion, motivation, concern, or personal experience. The labels come from a transformer-based classifier fine-tuned on about 900 hand-labeled student messages addressed to counselors.

A.4.0.2 Base model and architecture. We use `dccuchile/patana-chilean-spanish-bert` (Cañete et al., 2023), a 12-layer, 768-hidden-unit BERT model (Devlin et al., 2019) pre-trained on Chilean Spanish corpora that includes social-media and WhatsApp-like text. The architecture adds a dropout layer ($p = 0.2$) and a two-logit linear classification head to BERT’s pooled output. Following Sun et al. (2019), all parameters are fine-tuned, with AdamW (learning rate $2e-5$), batches of 16, five epochs, and inputs truncated at 128 tokens. Each label has its own binary cross-entropy loss. A label is assigned when its sigmoid probability passes 0.5, independently of the other label.

A.4.0.3 Evaluation. We split the labeled sample into a training set (80%) and a held-out test set (20%), stratifying on the joint distribution of the two labels. Hyperparameters are chosen by 5-fold stratified cross-validation within the training set. The test set is used only to compute the metrics reported below.

Table A-1: Patana multi-label classifier: Performance.

Label	Accuracy	Macro-F1
Factual	0.928	0.925
Subjective	0.940	0.530

A.4.0.4 Time to the first informative exchange. We define a student turn as *important* if it is predicted to be factual or subjective by the classifier above. For each session, the *time to the first informative exchange* is the number of minutes elapsed between the first student turn and the first substantive student turn. This captures how quickly the exchange moves past opening pleasantries toward informational content.

A.5 Semantic coherence via contextual embeddings

Building on recent text-as-data methods that represent language as high-dimensional vectors (Ash and Hansen, 2023), we measure coherence as the cosine similarity between dense representations of a prompt and its reply.

A.5.0.1 Pair construction. We go through each conversation’s turns in chronological order and form *input*→*output* pairs. If the same role has several consecutive turns, we first join them into one block, which is then paired with the next block from the other role. There is one set of pairs for each direction. *Counselor coherence* (user-to-system) pairs each student block with the counselor block that follows, so it measures how relevant the counselor’s reply is to what the student just said. *Subject coherence* (system-to-user) pairs a counselor block with the student block that follows it, and measures whether the student’s reply is topically aligned with the counselor’s prompt rather than a generic acknowledgement.

A.5.0.2 Embedding model. We encode every input and output with `intfloat/multilingual-e5-large` (Wang et al., 2024), a multilingual encoder trained on a large collection of weakly-supervised text pairs. We chose it because it does well on Spanish retrieval benchmarks and because it was trained for asymmetric retrieval with a `query:/passage:` prefix convention, which fits our prompt-and-reply pairs. Following that convention, we prepend `query:` to the input side and `passage:` to the output side. Embeddings are L_2 -normalized so that cosine similarity reduces to a dot product,

$$\text{cos_sim}(u, v) = \langle \tilde{e}_u, \tilde{e}_v \rangle, \quad \tilde{e}_x = e_x / \|e_x\|.$$

Cosine similarity is a standard measure of semantic relevance in information retrieval (?) and sentence-pair similarity (Reimers and Gurevych, 2019).

A.5.0.3 Coherence Examples: *Low coherence.*

Student: “I know it exists, but I don’t know the dates. I might be interested because I like UFRO, but I’m not familiar with the requirements.”

Counselor: “Have you filled out the FUAS?”

High coherence

Student: “I would like to know whether at the Universidad Alberto Hurtado (UAH) the degree in language teaching allows you to teach both at the primary and secondary levels.”

Counselor: “Yes, at the Universidad Alberto Hurtado (UAH), a degree in Language Teaching qualifies you to teach at both the primary and secondary levels. The program is designed to equip you with the skills needed to address the curriculum at both stages. This includes teaching methodologies, lesson planning, and assessment, allowing you to adapt to students’ needs at different stages of development. Would you like to know more about the graduation process or the curriculum at UAH?”

A.5.0.4 Engagement examples: *Low engagement.*

Counselor: “And of course, if you are applying through the FUAS, you may be eligible for tuition-free funding.”

Student: “How long does the early childhood education degree last, or does it depend on the university?”

High subject coherence.

Counselor: “Great! I’d like to ask you a couple of questions to better guide the conversation. Are you currently in your final year of high school or have you already graduated? Have you taken the PAES? And which teaching field are you interested in?”

Student: “I graduated this year, I already took the PAES, and I’m interested in pursuing a degree in Language and Communication teaching.”

A.5.0.5 Aggregation. For each conversation, we average the cosine similarity over all its pairs, separately for each direction. The two resulting measures, *counselor coherence* and *subject coherence*, are what the mechanisms section reports.

A.6 Interpersonal-quality scoring

The mechanisms section quantifies *motivation*, *informativeness*, and *empathy* through a hybrid human-in-the-loop pipeline. We also construct measures for *patience* and *beliefs*, but these are not part of the main mechanisms itemization. We follow recent work that uses semi-supervised ML to score open-ended text (Ash and Hansen, 2023; Gentzkow et al., 2019), and adapt it to WhatsApp counseling dialogues.

A.6.0.1 Rater recruitment and sessions. Rating took place in two structured sessions of about five hours each, on August 22 and August 29, 2025. We recruited 13 undergraduate psychology students from the Pontificia Universidad Católica de Chile for the task. Three identified as male and ten as female, and seven were in their fourth year of study and six in their third. Eleven of the thirteen had research experience. Conversations were shown with no indication of the arm, so raters were blind to treatment. Before scoring, they received standardized training on the instrument (the definitions of each dimension and sub-dimension, the four-point ordinal scale, and the not-applicable flag), and each session opened with a calibration exercise on anchor examples.

A.6.0.2 Human annotation. Raters scored a random sample of about 900 full conversations. There are ten ordinal dimensions in five constructs. *Patience* has two: tolerance of repeated questions, and no rushing or dismissive language. *Motivation* also has two (explicit encouragement toward teaching, general verbal support), as does *informativeness* (accuracy and completeness, clarity and pedagogical structure). *Beliefs* is a single dimension, whether the counselor elicits or addresses the student’s beliefs about teaching. *Empathy* has three: warmth and emotional acknowledgement, understanding of the student’s situation, and probing of the student’s concerns. Each is scored on a three-point scale. Following standard practice in content analysis (Krippendorff, 2018), a dimension can be marked *not applicable* if the conversation is too short or off-topic to judge. We use the median rating for each conversation and dimension (robust to a single outlying rater).

A.6.0.3 Feature construction. We then predict these labels for all conversations with a supervised classifier, using features built for counseling dialogues.

Role-decomposed BERT embeddings. We compute four `patana-chilean-spanish-bert` embeddings per conversation and concatenate them. They cover (i) the full text, with counselor and student turns interleaved; (ii) the counselor turns; (iii) the student turns; and (iv) the student’s last turn, which reflects the closing sentiment and any request left unresolved. We pool by taking the attention-masked mean of the last hidden layer. Reimers and Gurevych (2019) show that this does substantially better than [CLS] pooling on sentence-level tasks when the

encoder has no fine-tuned classification head. Empathy is mostly visible in the counselor’s words. Beliefs and motivation more often surface in what students disclose, which counselors may or may not pick up on. A single embedding of the whole conversation would dilute both signals, which is why we embed the roles separately.

Handcrafted counselor features. Next to the embeddings, we count specific behaviors in the counselor’s text with regex dictionaries taken from the literature on effective counseling communication. The counts cover volume and pacing (number of turns, turn lengths, counselor-to-user ratio), patience markers (*calma, tranquilo/a, paso a paso*), repetition markers (*de nuevo, otra vez*), motivation markers (*puedes, excelente, buen trabajo*) and exclamation-mark counts, informativeness markers (*porque, concepto, definición, explicar*) and question-mark counts, warmth markers (*entiendo, comprendo, me imagino*), understanding markers (*claro, veo, perfecto*), exploration markers (*qué piensas, por qué crees, qué opinas*) and an open-question pattern, example markers (*ejemplo*), logical connectors (*entonces, por lo tanto*), and lexical-complexity proxies (average word length, share of long words). The scaler that standardizes these counts is fit on the labeled subsample only and then applied to all conversations, to avoid information leakage. Concatenating the four embeddings with the standardized counts gives each conversation’s final feature vector.

A.6.0.4 Two-stage model. Two features of the labels complicate prediction. Quality dimensions are ordinal and correlated across constructs, and several are heavily imbalanced (the top category is rare). The model has two stages, a binary gate and, for conversations that pass it, one continuous regressor per conceptual group.

Stage 1: applicability gate. Some conversations are too thin to score. A Gradient Boosting Classifier (Friedman, 2001) identifies them, with *applicable* as the positive class. Conversations it labels not applicable get no Stage 2 score and are coded missing in the distributions we report.

Stage 2: composite regressor. For each conceptual group, a regression model predicts the conversation-level composite *directly*. We do not classify sub-dimensions one by one and average them. For group g and conversation i , the labeled composite is

$$\bar{y}_{ig} = \begin{cases} \frac{\sum_{d \in g: y_{id} \neq \text{missing}} y_{id}}{\sum_{d \in g} \mathbf{1}\{y_{id} \neq \text{missing}\}} & \text{if } \exists d \in g \text{ such that } y_{id} \neq \text{missing}, \\ \text{missing} & \text{otherwise.} \end{cases}$$

When some sub-dimensions are inapplicable or missing, the composite averages only the remaining ones. If all sub-dimensions in a group are missing, the composite is missing, and the conversation is dropped from that group’s training sample. We train the regressor on the labeled, non-missing $(\bar{y}_{ig}, x_i)$ pairs and predict $\hat{y}_{ig}$ for every conversation that the Stage 1 gate

flags as applicable. For each group we fit both a Random Forest regressor (Breiman, 2001) and a Gradient Boosting regressor (Friedman, 2001), and keep the family and hyperparameters with the lowest five-fold cross-validated MAE on the training fold, using folds stratified on tertiles of the composite.

A.6.0.5 Model selection and validation. Validation works the same way in both stages. An 80/20 split of the labeled sample, stratified on the applicability label, sets aside a test set that is never used for training or model selection. Inside the 80% fold, we run five-fold stratified cross-validation with a grid search. The gate is scored on macro-F1, which penalizes ignoring the minority class, and the Stage 2 regressors on MAE, the prediction error in the units of the composite. The gate’s grid covers the number of estimators, maximum depth, and learning rate of a Gradient Boosting classifier. In Stage 2 we run two grid searches per group, one over Random Forest regressors (number of trees, maximum depth) and one over Gradient Boosting regressors (number of trees, maximum depth, learning rate), and keep the winner. The selected configurations are evaluated once on the 20% test set. The predictions in the main text come from refitting them on the whole 80% fold and applying them to every conversation the Stage 1 gate flags as applicable.

A.6.0.6 Applicability-gate performance. Cross-validation selects the Gradient Boosting hyperparameters $\hat{n}_{\text{est}} = 300$, $\hat{d}_{\text{max}} = 3$, and $\hat{\eta} = 0.1$, with a five-fold CV macro-F1 of 0.883. Table A-2 reports the gate’s performance on the 20% held-out test set. Of the 9,327 conversations in the full sample, the gate flags 3,140 (33.7%) as applicable for Stage-2 quality scoring.

Table A-2: Applicability gate: held-out performance.

	Accuracy	Macro-F1
Applicability gate	0.886	0.885

A.6.0.7 Per-group Stage-2 performance. For each conceptual group, Table A-3 reports the model family selected by cross-validation and its MAE, RMSE, R^2 , Pearson r , and Spearman ρ on the 20% held-out set. Table A-4 groups predicted and true composites into *Low* (< 1.5), *Medium* (1.5–2.5), and *High* (> 2.5), and reports how often the prediction falls in the right bucket or within one bucket of it. Motivation, informativeness, and empathy, the groups with several dimensions, reach Pearson correlations of 0.6 to 0.8 with the human labels, with correspondingly high bucket accuracy. Patience and beliefs do worse. Patience barely varies across the labeled conversations, leaving the regressor little to work with, and beliefs is

almost binary (very few conversations explicitly elicit student beliefs), so a continuous model is unnecessary. We treat both as auxiliary measures.

Table A-3: Stage-2 composite-regressor performance on the held-out (20%) test set.

Group	Family	MAE	RMSE	R^2	Pearson r	Spearman ρ
Patience	GBT	0.366	0.476	-0.143	0.083	0.184
Motivation	RF	0.441	0.536	0.386	0.651	0.527
Informativeness	GBT	0.384	0.465	0.407	0.638	0.556
Beliefs	RF	0.072	0.142	-0.127	-0.037	0.115
Empathy	RF	0.459	0.569	0.551	0.802	0.668

Table A-4: Composite-regressor bucket accuracy — held-out (20%) test set.

Group	Exact bucket	Within ± 1 bucket
Patience	0.652	0.971
Motivation	0.371	1.000
Informativeness	0.732	0.986
Beliefs	1.000	1.000
Empathy	0.719	1.000

From composites to indices. The *empathy*, *motivation*, and *informativeness* indices in the mechanisms section are the predictions of the corresponding Stage-2 regressors, one per group and conversation, with no further transformation. Conversations the Stage-1 gate flags as not applicable are missing, and arm-level means cover only conversations where the construct is well defined.

Non-applicable sub-dimensions are likewise left out of the composite targets. A conversation that scores positively on one sub-dimension and is missing another is scored on the first alone, instead of being pulled toward the middle of the scale.

Whether a sub-dimension is not applicable depends on where the student takes the conversation and on the opportunities that come up in it. Inapplicability thus reflects the context of the interaction and says nothing about the counselor’s performance. We exclude these cases from the aggregation so that counselors are not penalized on dimensions they had no chance to show.

A.7 Topic classification

A.7.0.1 Taxonomy construction. The topic taxonomy was constructed following an inductive, AI-assisted thematic analysis approach [Naeem et al. \(2025\)](#). Conversations were first

anonymized by removing any information that could identify the student or the human counselor, including names, gender, age, and place of residence. We then used Claude as an inductive coding assistant to identify recurrent themes in the conversation logs. We asked the model to propose candidate themes based on a random sample of conversations, merge overlapping ones, and group them into macro-categories consistent with Elige Educar’s program guidelines. We used the model only for this first pass over a large corpus. The resulting taxonomy then went through several rounds of human review, revision, and validation. In this process, the proposed categories were reviewed against the conversation logs and the program’s conceptual framework, merging two pairs of overlapping themes and splitting one category that conflated distinct student concerns. The final taxonomy comprises 18 subcategories nested within three macro-categories: (1) *admissions and application process* (application process, entry requirements, scholarships and financial aid, university inquiries, administrative inquiries, PAES concerns, and the Teaching Vocation Scholarship (BVP)); (2) *career and professional outlook* (the teaching degree, labor market prospects, curricula, professional internships, accreditation and program quality, compatibility with work, academic demands, and other careers); and (3) *vocational guidance* (personal experiences, vocational orientation, and conversational interactions).

A.7.0.2 Automated topic classification. Conversations were classified into the 18 subcategories using the ChatGPT API (model `gpt-4o-mini-2024-07-18`), with temperature set to 0 and a fixed seed (42) to ensure replicability. We classified each message and then aggregated to the conversation level, coding a subcategory as 1 if at least one student message in the conversation addresses it.

The prompt told the model to classify only what students wrote and to ignore the replies of counselors and the chatbot. Without this instruction, the model tended to classify the counselor’s replies. For instance, if a student asked about scholarships in general and the counselor answered with details on the Teaching Vocation Scholarship (BVP), the model would wrongly record that the student was interested in the BVP. The final prompt therefore states this rule prominently at the start. It also includes six examples from this setting of correct and incorrect classification, with an emphasis on hierarchically related categories, a reference example placed right before each conversation, and sharper definitions for pairs of subcategories with similar meanings.

We revised the prompt over seven cycles. In each one, we classified a pilot set of 500 conversations, read the output to find systematic errors, changed the prompt to fix them, and ran the pilot again. The counselor-versus-student error described above came up in the first cycle. Later cycles dealt with over-classification (we flagged any conversation assigned ten or more subcategories) and with logical inconsistencies between hierarchically related categories, for example a conversation coded for the BVP but not for its parent category, scholarships and financial aid. After each revision, and before running the classifier on all conversations, an

automated function checked the output for critical errors and computed distributional health metrics.

A.7.0.3 Classification prompt. The final classification prompt used in the automated pipeline is reproduced below.⁴

Listing 1: Topic classification prompt (final version)

Eres un clasificador de mensajes educativos.

REGLA DE ORO: Clasifica SOLO lo que el ESTUDIANTE/USUARIO pregunta, menciona o dice. NUNCA clasifiques lo que el counselor/CHATBOT/KAI responde o explica.

== PROCESO ==

PASO 1: Identificar si la conversacion trata sobre pedagogia
La decision se basa UNICAMENTE en los mensajes del ESTUDIANTE.
Ignora completamente lo que dice el counselor/Kai.

es_sobre_pedagogia = 1 SOLO si el estudiante hace al menos UNA de estas cosas:

- Escribe texto natural confirmando interes ("si", "me interesa", "quiero estudiar pedagogia", etc.)
- Menciona explicitamente una carrera de pedagogia con sus propias palabras
- Hace una pregunta sobre pedagogia, universidades, becas, postulacion, etc.

es_sobre_pedagogia = 0 en TODOS estos casos:

- El estudiante no envia ningun mensaje de texto (solo selecciona botones sin contenido propio)
- El estudiante no responde nada
- Solo el counselor/Kai habla sobre pedagogia, pero el estudiante no confirma ni pregunta nada
- El estudiante responde con saludos unicamente ("hola", "gracias") sin referencia a pedagogia

⁴The word "*counselor*" in the prompt appears exactly as in the version we ran, and matches the term used in the rest of the paper.

REGLA CLAVE: Si no hay mensajes de texto del estudiante que demuestren interes o consulta activa → es_sobre_pedagogia = 0

PASO 2: Extraer SOLO los mensajes del ESTUDIANTE/USUARIO
Ignora completamente los mensajes del counselor/KAI.

PASO 3: Clasificar cada mensaje del ESTUDIANTE en las categorias que apliquen

== CATEGORIAS ==

POSTULACION Y ADMISION:

1. proceso_postulacion: Estudiante pregunta sobre como/cuando postular
2. requisitos_ingreso: Estudiante pregunta sobre puntajes , NEM, ranking
3. becas_beneficios: Estudiante pregunta "becas", "financiamiento", "gratuidad" (EN GENERAL)
4. consultas_universidades: Estudiante pregunta "que universidad", "cual es mejor"
5. consultas_administrativas: Estudiante pregunta sobre matricula , documentos
6. inquietudes_paes: Estudiante pregunta sobre PAES, fechas , resultados
7. beca_vocacion_profesor: Estudiante pregunta ESPECIFICAMENTE "BVP" o "Beca Vocacion de Profesor"

CARRERA Y PROYECCION:

8. carrera_pedagogia: Estudiante menciona o pregunta sobre pedagogia , especialidades
9. campo_laboral: Estudiante pregunta sobre trabajo , sueldo , empleabilidad
10. mallas_curriculares: Estudiante pregunta sobre ramos , asignaturas , plan de estudios
11. practicas_profesionales: Estudiante pregunta sobre practicas
12. acreditacion_calidad: Estudiante pregunta sobre

- acreditacion , calidad
13. compatibilidad_trabajo: Estudiante pregunta si puede estudiar y trabajar
 14. exigencia_academica: Estudiante pregunta si es dificil , carga academica
 15. otras_carreras: Estudiante menciona otras carreras

ORIENTACION VOCACIONAL:

16. experiencias_personales: Estudiante expresa motivacion , gusto o vinculo personal con la docencia. Incluye frases directas ("me gusta enseñar", "quiero ser profesor") Y tambien relatos mas narrativos donde el estudiante describe por que le atrae la carrera , experiencias propias con ninos/jovenes , o razones personales para elegir pedagogia
17. orientacion_vocacional: Estudiante expresa dudas sobre que estudiar
18. interacciones_conversacionales: Solo saludos , gracias , despedidas SIN contenido

== EJEMPLOS CRITICOS ==

EJEMPLO 1 – counselor HABLA, ESTUDIANTE NO RESPONDE:

Kai: "Hola , vimos que tienes interes en pedagogia."

Estudiante: [no responde / selecciona boton sin texto]

→ es_sobre_pedagogia = 0

→ NO clasificar ningun mensaje

EJEMPLO 2 – counselor HABLA, ESTUDIANTE SOLO SALUDA:

Kai: "Hola , te interesa estudiar pedagogia?"

Estudiante: "Hola"

→ es_sobre_pedagogia = 0

→ cat_18_interacciones_conversacionales = 1

EJEMPLO 3 – ESTUDIANTE CONFIRMA CON TEXTO:

counselor: "Te interesa pedagogia?"

Estudiante: "Si , me interesa pedagogia basica"

→ es_sobre_pedagogia = 1

→ cat_8_carrera_pedagogia = 1 ,

cat_16_experiencias_personales = 1

EJEMPLO 3b – cat_16 NARRATIVO:

Estudiante: "Me gusta la idea de trabajar con esos niños,
entender por que son así"

→ cat_16_experiencias_personales = 1

EJEMPLO 4 – BECAS (CORRECTO):

Estudiante: "Becas"

counselor: [Explica sobre FUAS, gratuidad, BVP extensamente]

→ cat_3_becas_beneficios = 1

→ cat_7_beca_vocacion_profesor = 0

EJEMPLO 5 – BVP ESPECIFICA (CORRECTO):

Estudiante: "Quiero saber sobre la Beca Vocacion de Profesor"

→ cat_3_becas_beneficios = 1,

cat_7_beca_vocacion_profesor = 1

EJEMPLO 6 – RESPUESTA CORTA CON CONTEXTO:

counselor: "Te interesa pedagogia en historia o matematicas?"

Estudiante: "La primera"

→ es_sobre_pedagogia = 1

→ cat_8_carrera_pedagogia = 1

== FORMATO DE RESPUESTA ==

Responde SOLO con JSON (sin “‘json ni markdown):

```
{
  "es_sobre_pedagogia": 1,
  "mensajes": [
    {
      "mensaje_id": "msg-1",
      "contenido": "texto del mensaje DEL ESTUDIANTE",
      "categorias": {
        "cat_1_proceso_postulacion": 0,
        "cat_2_requisitos_ingreso": 0,
        "cat_3_becas_beneficios": 1,
```

```

    "cat_4_consultas_universidades": 0,
    "cat_5_consultas_administrativas": 0,
    "cat_6_inquietudes_paes": 0,
    "cat_7_beca_vocacion_profesor": 0,
    "cat_8_carrera_pedagogia": 0,
    "cat_9_campo_laboral": 0,
    "cat_10_mallas_curriculares": 0,
    "cat_11_practicas_profesionales": 0,
    "cat_12_acreditacion_calidad": 0,
    "cat_13_compatibilidad_trabajo": 0,
    "cat_14_exigencia_academica": 0,
    "cat_15_otras_carreras": 0,
    "cat_16_experiencias_personales": 0,
    "cat_17_orientacion_vocacional": 0,
    "cat_18_interacciones_conversacionales": 0
  },
  "total_temas": 1
}
]
}

```

RECUERDA: Clasifica SOLO lo que el ESTUDIANTE dice,
NO lo que el counselor explica.