Generative AI in Financial Services: A Field Experiment in Insurance Advisory*

Job Market Paper

Yingju Ma

University of Cambridge

November 1, 2026

Click here for the latest version

Abstract

I study a generative-AI tool that offers insurance brokers real-time reply suggestions, using two randomized phases of its rollout, involving 1,070 licensed brokers at China's largest insurance brokerage. In a ten-week pilot, 239 randomly selected brokers were given access to the AI tool; in a subsequent 20-day experiment, about 165,000 customers were randomly assigned to conversations in which the broker either had or did not have access to AI advice. Access to AI advice raises realized premium by 28 percent within 15 days among customers of non-pilot brokers, and customers buy more life insurance. AI's benefits depend on broker ability and the product. In life insurance and large policies, where premiums per sale are higher, gains rise with brokers' baseline ability; in health insurance, the tool narrows performance gaps across brokers. Brokers who adopt suggestions more often do not gain more, suggesting that adoption frequency is an imperfect proxy for AI's value. Pilot brokers' advantage emerges after the initial weeks and persists after full roll-out, even after adjusting for their greater effort, consistent with learning.

Keywords: Generative AI; Financial Advice; Life Insurance; Field Experiment; Skill Complementarity; Human-AI Collaboration. **JEL codes:** D83, G22, G51, J24, O33

*I am grateful to Nikhil Agarwal, Vivek Bhattacharya, Stefan Bucher, Gilles Chemla, Taha Choukhmane, Sara Fisher Ellison, Eric French, Xavier Giroud, Sanjeev Goyal, Amy Handlan, David Hirshleifer, Kai Liu, Gordon Phillips, Tobias Salz, Paul Scott, Robert Townsend, Benjamin Vatter, and Constantine Yannelis for helpful comments and discussions, to Ziang Li and Edison Yu for discussing the paper, and to seminar and conference participants at the MIT IO Lunch, the University of Cambridge, the Imperial College Fintech Conference, the 2026 RAPS/RCFS Europe Conference, and the Philadelphia Fed Frontiers in Machine Learning and Economics Conference. The author acknowledges and appreciates the support from the *Digital Economy Open Research Platform* (www.deor.org.cn). All data was sampled and desensitized, and was remotely analyzed on the *Ant Open Research Laboratory* in an Ant Group Environment which is only remotely accessible for empirical analysis. The views expressed are solely my own and do not represent the views of Ant Group or Ant Insurance.

1 Introduction

Households making consequential financial decisions often rely on intermediaries to acquire information, interpret complex products and turn that information into advice they can trust. Generative AI is entering this work: financial institutions are deploying it in customer service and advice, and supervisors ask that customer-facing uses with material consequences remain under effective human oversight (Financial Stability Board, 2026). In insurance, a broker must understand a customer's circumstances, explain a contract and help the customer act on the advice.[1] A generative-AI tool can draw on product knowledge and customer history to draft a tailored reply, while the human intermediary remains responsible for what reaches the customer.[2] Such a tool could spread expertise to less capable intermediaries or raise the return to the expertise they already have; which of the two happens is central to understanding the role of human skill as AI enters professional services.

This paper shows that when brokers receive generative-AI reply suggestions, customers buy more life insurance and pay more premium; higher-ability brokers gain more with customers who already know them, mainly through large life policies; and how often brokers adopt suggestions is a poor guide to these gains.

I study a production-grade generative-AI tool that offers licensed insurance brokers real-time reply suggestions during live advisory conversations at Ant Insurance, China's largest insurance brokerage. The broker can send, edit or ignore each AI suggestion. The tool was rolled out in three phases: a ten-week randomized broker-level pilot, a 20-day customer-level experiment, and full rollout. In the pilot, 253 randomly selected brokers had early access to the AI tool for all of their customers. In the subsequent experiment, every broker had the tool, and each customer was randomly assigned to the Treat or Control group at the first message, when the tool switches on. About 80 percent of customers were assigned to Treat. Brokers served both groups, knew each customer's group and could use the tool only for Treat customers. The estimand is the effect of access to AI advice: the broker's opportunity to use AI suggestions in a given conversation. Because Control assignment withdrew a tool that pilot brokers had used for ten weeks, the main sample contains the 124,347 customers of the 831 non-pilot brokers; platform records follow their later conversations and purchases over the whole rollout.

I report four main findings. First, *access to AI advice increases life-insurance purchases and realized premium* (Section 3). Premium per customer rises by 88 yuan within 15 days, 28 percent of the Control mean of 315 yuan, and by 152 yuan within 30 days, or

[1] Sales incentives also shape this interaction: commission-based compensation can encourage unsuitable sales (Inderst and Ottaviani, 2012; Anagol et al., 2017; Egan, 2019).

[2] Salvi et al. (2025) provide experimental evidence that generative AI can produce persuasive, personalized arguments.

29 percent. The probability of buying life insurance rises by 0.33 percentage points within 15 days and 0.47 points within 30 days, about 14 and 17 percent of the Control rates. Randomization inference gives p-values of 0.002 for the life-insurance effect and 0.025 for 30-day premium; the 15-day premium effect is weaker under this procedure ($p = 0.064$) and significant under a wild cluster bootstrap with adjustment for the three main tests (Holm $p = 0.044$). Poisson estimates with broker fixed effects give the same 28 percent, with a Holm-adjusted p-value of 0.066 for each of the three tests. The overall purchase effect is not statistically significant within 15 days and is 0.28 percentage points within 30 days.

Second, and central to the paper, *the gains of higher-ability brokers depend on the product and on whether the customer already knows the broker* (Section 4). Ability is an index of brokers' sales performance (premium, conversions and conversion rate) and tenure in the ten pilot weeks before the experiment. On average, each standard deviation of the index raises the 15-day premium effect by 107 yuan, about a third of the Control mean (Holm-adjusted $p = 0.030$), and the 30-day effect by 155 yuan. The gradient is concentrated in life insurance, in large policies and among returning non-buyers, customers who had visited or chatted with a broker but had not bought, and within them among those who had already chatted with the same broker: for these customers the premium gradient is 189 yuan per standard deviation, against an insignificant -59 yuan for the other returning non-buyers, and it works mainly through life policies of 30,000 yuan or more. For new customers the effect varies little with ability, and in health insurance it is, if anything, larger for weaker brokers; the life-insurance gradient exceeds the health-insurance gradient by 0.33 percentage points per standard deviation (Holm-adjusted $p = 0.044$). Among returning non-buyers, about 70 percent of the premium gap between the top and bottom ability quintiles comes from larger purchases per buyer. Because brokers know which customers have the tool, I read the premium gradient as an upper bound on what the tool adds to stronger brokers' premium at full deployment (Section 4.4). The product contrast has a second reading: an average life policy pays the broker more commission per sale than a health policy, so the pattern reflects what a sale is worth to the broker as well as what the advice requires.

Tenure alone gives a similar gradient of 95 yuan. By tenure group, the premium effect is 23 yuan for brokers with less than three months of tenure and 236 yuan for brokers with more than a year; the trend across four tenure groups is significant at the 10 percent level, although equality across the groups is not rejected. In customer support, AI lets new workers perform like experienced ones (Brynjolfsson et al., 2025); in insurance advice the newest brokers gain least.

Third, *adoption is not value* (Section 5). Because each broker serves both groups, I estimate each broker's premium effect from one half of the customers and the broker's adoption share from the other half, so that the two measures share no sampling noise.

Brokers who adopt a larger share of the AI suggestions they receive have smaller premium gains (corrected correlation -0.39), while the gains correlate positively with ability (0.42). Over the three weeks, adoption rises sharply while the premium effect falls (the weekly differences are not statistically significant). Strong brokers gain most without adopting most. Brokers also lean on the tool when they are busy: on days when a broker happens to receive more Treat customers, she adopts a larger share of suggestions, all of it by sending them with parts deleted rather than editing them; on busier days the gain per Treat customer is also smaller.

Fourth, *the effects persist, for customers and for brokers* (Sections 3 and 6). For customers, access raises the probability that a customer returns to chat on days 15 to 45 by 1.37 percentage points, about 9 percent of the Control rate, and the probability of another purchase by day 45 by 0.13 points, about 20 percent; both survive adjustment for multiple outcomes. For brokers, the evidence comes from the pilot: conditional on rank and on pre-pilot sales and effort, pilot brokers' advantage in daily premium is close to zero in the first five pilot weeks, appears from week 6 and persists at roughly a fifth of the non-pilot mean after every broker has the tool. The timing of its appearance coincides with rising demand for savings-type life insurance, as guaranteed rates were expected to fall, so the onset alone is not evidence of learning. The persistence is: a pure access effect would vanish once all brokers have the tool, so the persistent advantage is consistent with learning. It is not a matter of working more: after adjustment, pilot brokers record only a few percent more work, and their premium per work day is 13 percent higher, mostly through larger purchases. Pilot brokers, however, already worked more before the pilot (joint balance $p = 0.0001$), so learning and residual selection on unobserved effort or skill cannot be separated.

To interpret why the gains differ by product and by the customer's history with the broker, I use a model in which the tool supplies content and the customer weighs it by how far she trusts the broker (Section 7). In health insurance, whose terms customers can check, content is enough and the tool narrows gaps across brokers. In savings-type life insurance, whose suitability is revealed only over years, content counts only as far as the customer trusts the broker, so the tool is worth more to stronger brokers with customers who already know them. The model interprets these patterns; it does not measure credibility.

These results contribute in three ways. First, they reconcile, within one experiment, evidence that AI levels performance in customer support, writing and programming (Brynjolfsson et al., 2025; Noy and Zhang, 2023; Cui et al., 2026) with evidence that it amplifies performance among entrepreneurs (Otis et al., 2026): holding the tool, the brokers, the customers and the days fixed, the product and the customer's relationship with the broker decide whether AI substitutes for or complements ability. Second, in advice the tool appears to steepen rather than flatten the experience curve. Third, the

paper links suggestion-level adoption records to randomized revenue effects and shows that adoption frequency misstates where value is created. Since firms and researchers often measure AI by its use (Babina, 2026; Humlum and Vestergaard, 2025b), the lesson is general. For organizations, value should be judged by realized customer outcomes rather than adoption counts, and a tool that raises the return to expertise can widen performance differences within an occupation unless training and customer allocation account for it.

Related literature. Research on *AI, worker skill and the organization of production* asks who gains from the technology. Generative AI produces new content and reaches a far broader set of cognitive tasks than predictive machine learning (Eloundou et al., 2024). Brynjolfsson et al. (2025) study a staggered deployment in customer support, and randomized experiments examine writing, coding and consulting (Noy and Zhang, 2023; Peng et al., 2023; Cui et al., 2026; Dell'Acqua et al., 2026). Several find larger gains among less-experienced or lower-performing workers, as does work on predictive AI in taxi navigation (Kanazawa et al., 2026). Dell'Acqua et al. (2026) show that gains depend on whether a task lies inside the technology's capabilities, and Otis et al. (2026) find that initially higher-performing entrepreneurs benefit more, with implementation choices helping explain the difference. These patterns speak to how technology changes the value of human capabilities and the allocation of tasks (Acemoglu and Restrepo, 2018, 2019; Autor, 2024; Hoffmann et al., 2024) and to the organization of knowledge (Garicano, 2000; Ferreira and Li, 2026). I show, within one deployment, that the task determines whether the same tool amplifies skill.

Research on *financial advice and intermediaries* studies how information, incentives and trust shape advice. Models examine conflicts between information provision and sales incentives (Inderst and Ottaviani, 2012; Gennaioli et al., 2015); empirical research studies steering, misconduct and brokers' own beliefs.[3] Insurance research studies how information and experience affect demand and the match between holdings and household needs (Koijen et al., 2016; Cai et al., 2015, 2020). Predictive AI uses digital footprints to assess default risk (Berg et al., 2020) and changes the distribution of credit-market gains (Fuster et al., 2022). Closest to my setting, Liu (2026) gives insurance brokers purchase-intent scores based on customers' digital footprints; sales rise, but attention can shift away from risk-relevant information. Robo-advising research examines diversification, fees, access, personalization and the value of human contact (D'Acunto

[3]Studies document steering toward costly or unsuitable products (Bergstresser et al., 2009; Christoffersen et al., 2013; Mullainathan et al., 2012; Anagol et al., 2017; Egan, 2019), misconduct (Egan et al., 2019), limited customization (Foerster et al., 2017), and the influence of intermediaries' own beliefs (Linnainmaa et al., 2021). Research on regulation and fiduciary obligations examines how institutional constraints shape advice quality and market participation (Charoenwong et al., 2019; Egan et al., 2022; Bhattacharya et al., 2025).

et al., 2019; Rossi and Utkus, 2024; Reher and Sokolinski, 2024; Jiang et al., 2025; Greig et al., 2025). With generative AI, Fieberg et al. (2025) evaluate LLM portfolio advice across investor profiles, and Choukhmane et al. (2026) study household prompts and the lifetime implications of following LLM advice. On a large Chinese fintech platform, He et al. (2026) show that mutual fund investors follow a generative-AI chatbot's suggestions in about half of interactions and that those who follow them trade less and earn higher returns, while Yang et al. (2026) study human bankers with final authority over AI-based investment advice.[4] In most of this work, AI advises households directly or informs a single decision; my setting places a licensed intermediary between the AI and the customer, links randomized access to brokers' use and realized transactions, and shows that the value of AI in advice depends on the intermediary's expertise.

Research on *human use of algorithmic advice* studies when people follow algorithms. Machine predictions create value through the human decisions they inform (Kleinberg et al., 2018; Agrawal et al., 2019). Responses to algorithms depend on prior errors and expertise (Dietvorst et al., 2015, 2018; Logg et al., 2019). Expert assistance can be limited by under-response and effort crowd-out (Agarwal et al., 2023, 2025), and Angelova et al. (2026) show that the value of discretionary overrides differs sharply across decision-makers. Combining humans and AI need not outperform the better of the two (Vaccaro et al., 2024). In my setting, the finding that brokers who adopt more gain less, while abler brokers gain more, is consistent with selective use as a source of complementarity, although receipt of a suggestion also depends on whether a conversation lasts long enough for the tool to generate one. I discuss research on learning and technology adoption with the pilot evidence in Section 6.

2 Institutional Setting, Experimental Design, and Data

2.1 Ant Insurance and the Advisory Process

I study licensed insurance brokers at Ant Insurance in China. China's insurance industry collected 5.7 trillion RMB in premiums in 2024, of which about 17 percent was intermediated by specialized agencies and brokerages; Ant Insurance has ranked first among all intermediaries since 2022, with about 6 to 7 percent of the intermediated premium. It connects over 600 million customers who have purchased insurance through the Alipay app with products from more than 90 partner insurers.[5] I focus on life and health insurance, which account for 90 percent of the brokers' policies and 95 percent of their

[4] For reviews, see Goldstein et al. (2019) on fintech and Eisfeldt and Schubert (2025) on generative AI and finance. In another intermediary setting, Cong et al. (2025) study AI support for real-estate brokers in China.

[5] Source: *Yearbook of China's Insurance* (2025). Appendix A gives further detail on the market and the product lines.

premium. Life insurance, with a median policy premium of 10,000 yuan, accounts for almost all of the premium; health insurance, which in the records covers both critical-illness and medical insurance, has a median premium of 486 yuan and is usually bought sooner after the first chat (Appendix Table A1).

Each customer who seeks advice is randomly matched to a licensed broker by the platform. The broker is the only party who writes to the customer: the customer never interacts with the generative-AI tool directly. The conversation runs through the platform's instant-messaging (IM) interface and often spans several days; the broker may also follow up by telephone, which the tool does not mediate (Appendix Figure A2). A customer's path has three steps. A *visit* is a tap on a broker's avatar, which opens the chat window and triggers preset replies; a *chat* begins when the customer sends a message; a *purchase* is a policy bought by the customer. The records hold only policies that were not refunded, so purchases and premiums, the total amount the customer paid, measure realized revenue. Purchases occur almost entirely among customers who chat.

Each broker carries a five-step seniority rank (L1 to L5), a coarse but public signal of broker quality; among the non-pilot brokers, whose customers form the main sample (Section 2.4), 56.4 percent are L2. Median tenure at the start of the pilot is about seven months. A broker serves many customers in parallel, starting first conversations with about seven to eight new customers a day.

2.2 The Generative-AI Tool

The tool combines a fine-tuned large language model, an insurance knowledge base and a corpus of past advisory transcripts (Appendix Figure A1). It starts working only after the customer's first message. For each incoming message, it conditions on the message, the customer's past conversation history and personal information (age, gender, city, family composition, past insurance activity) and generates suggested replies for the broker, who decides whether to use a suggestion, modify it, or write an independent reply; the customer does not see the tool. Generation takes time, so a short conversation can end before a suggestion arrives, and zero exposure can reflect a short conversation rather than a broker's choice.

The platform records how the broker handles each suggestion in one of seven states (Appendix A). I call a suggestion the broker received an *exposure* and a suggestion the broker accepted in any form an *adoption*. Neither is a customer action, and receipt alone does not establish that the broker read the suggestion. The broker adopted at least one suggestion for about one in five Treat customers; Section 5 documents how adoption varies across brokers. The tool did not change during the experiment (no model updates, prompt changes or changes to the quality filter), and the pilot used the

same tool.

2.3 Phased Rollout and Randomization

The tool was rolled out in three phases (Figure 1): a randomized broker-level pilot of ten weeks; a customer-level experiment of 20 days, in which every broker had the tool and customers were randomly assigned to the Treat or Control group; and full rollout to all customers. I count days from the first day of the experiment: Day 1 is its first day and Day 20 its last.

The pilot. In all, 253 pilot brokers had ten weeks of early access to the tool for every customer they served; 239 of them serve customers in the experiment. They were drawn at random within rank, but the selection share differed by rank, and pilot brokers had worked more than other brokers of the same rank just before the pilot. Three broker frames recur. The main pilot frame contains the brokers of ranks L2 to L5 active on at least three of the nine days before the pilot: 862 brokers, of whom 187 are pilot brokers. The roster frame (1,074 brokers of ranks L2 to L5 at the start of the pilot, 204 pilot brokers) and the first-pilot-week frame (858 brokers active on at least three days of the first pilot week, 187 pilot brokers) serve as comparisons. Appendix Table A36 gives selection by rank. In the main frame, pilot and non-pilot brokers sold similar amounts and had similar tenure before the pilot but spent more time on the telephone and in IM and were active on more days (joint test of seven baseline covariates, $p = 0.0001$). Every pilot comparison therefore conditions on rank and on pre-pilot outcomes and covariates, and inference uses permutations of the pilot indicator within rank. The nine days before the pilot are used only for this frame, its balance test and these covariates.

The experiment. Every broker received the tool two days before Day 1. I call these the transition days and exclude them from the experiment window; customers whose first chat fell on them stay in the sample with that day as their first-chat date. For 20 days each customer was randomly assigned, at the first message, to the Treat group, in which the tool was on in the broker's workspace for that conversation, or to the Control group, in which it was off. The same broker typically handled both groups each day (Figure 3).

Full rollout. After Day 20 the tool was on for all customers. This period has no Control group; I use its first seven weeks to follow experiment customers.

Assignment at the first message. Because the group is assigned at the customer's first message and the tool starts only after that message, whether a customer writes cannot

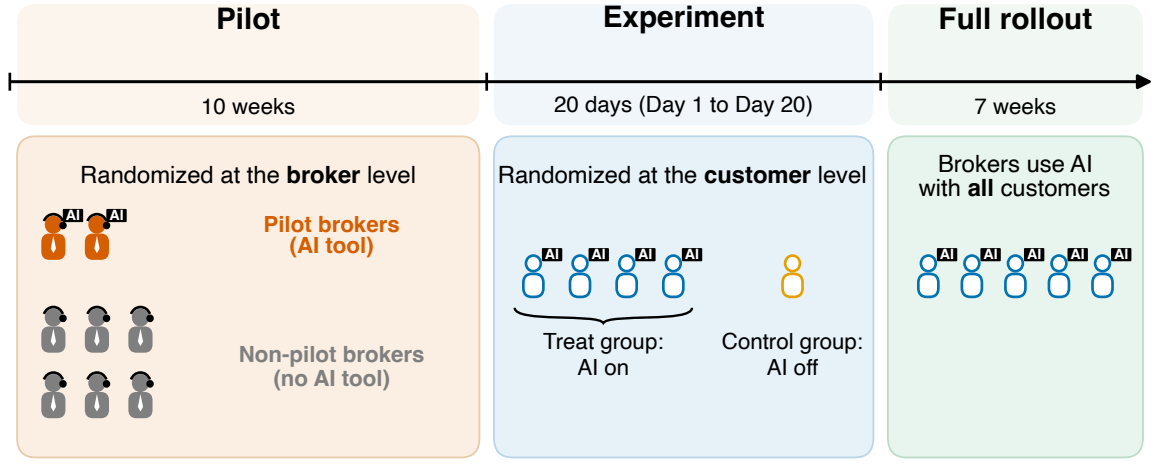


Figure 1: The Phased Rollout of the Generative-AI Tool

Notes: The three phases of the rollout (not to scale). Figures with headsets are brokers; outlined figures are customers; the AI label marks conversations in which the broker receives AI suggestions. In the pilot, 253 pilot brokers, drawn at random within rank with selection shares that differed by rank, used the tool with all of their customers. In the experiment, every broker had the tool and each customer was assigned at the first message to the Treat group (AI suggestions on) or the Control group (AI suggestions off). In full rollout the tool was on for all customers. Icon counts show proportions only: pilot brokers are about one in five brokers, and about four in five customers in the experiment are in the Treat group.

respond to the assignment; the experimental frame is therefore the set of customers who chat. I take each customer's group from the record of the first chat, the first day with a recorded dialogue round. The Treat share is about 0.80 but varies by day: among customers of non-pilot brokers it is 0.68 on Day 1 and between 0.80 and 0.82 on the other 19 days (Figure 2(b)). I do not rely on a fixed nominal probability; every Treat and Control comparison conditions on broker and first-chat-date fixed effects.

A Treat assignment *enables* AI in the broker's workspace but does not *force* it: among customers of non-pilot brokers (the main sample; Section 2.4), the broker received at least one suggestion for 45.6 percent of Treat customers and 0.7 percent of Control customers. I treat actual use as endogenous and instrument it with assignment in Section 5. Brokers were not blind to the assignment: serving a Control customer, the broker had no access to AI advice.

Identification assumptions. Random assignment within broker and first-chat date identifies the effect of access to AI advice (Section 2.6). Reading that effect as operating only through received suggestions requires an exclusion restriction, used for the IV estimates in Section 5: brokers know whether they have access to AI advice, and this may affect their effort or telephone follow-up. Cross-customer spillovers also separate the customer-level contrast from full deployment, since knowledge gained with Treat customers may carry into Control conversations. Sections 3 and 6 examine these channels.

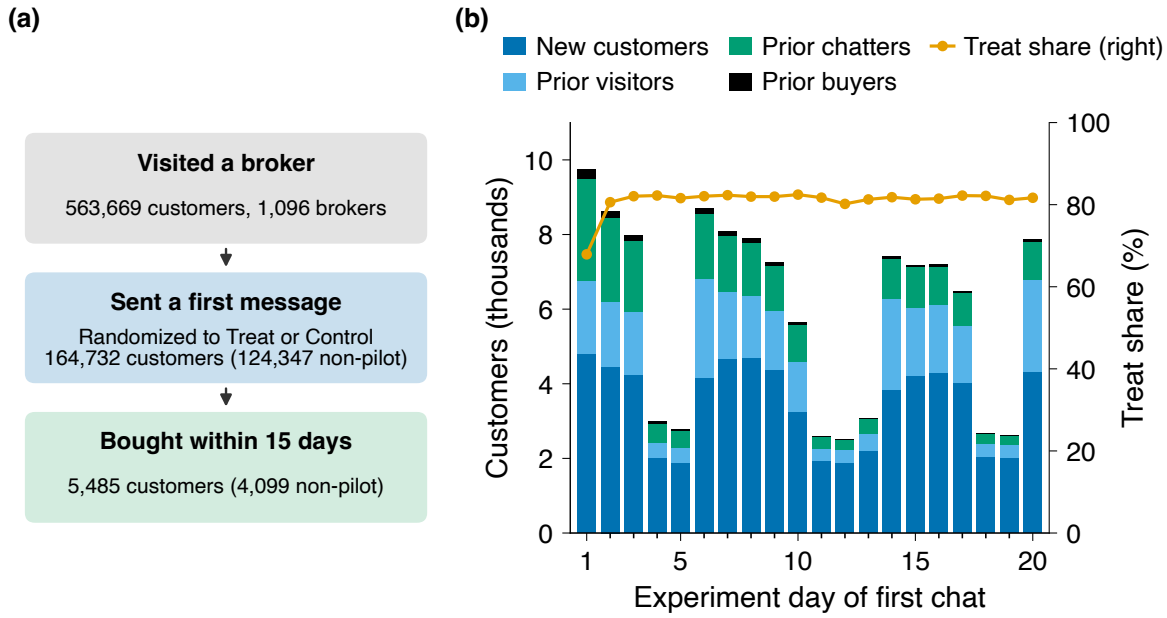


Figure 2: First Chats in the Experiment

Notes: Panel (a) shows the funnel: customers who visit a broker during the experiment, customers who chat (the tool switches on and the customer is randomly assigned at the first message), and customers who buy within 15 days. Panel (b) shows customers of non-pilot brokers by the experiment day of their first chat, stacked by prior funnel stage (deepest stage reached with any broker in the ten weeks before the first chat), with the Treat share of the day on the right axis. The Treat share is about 0.80 and varies by day (lower on Day 1); first-chat-date fixed effects absorb this variation. Days 11 to 13 include a public holiday, when fewer customers start chatting.

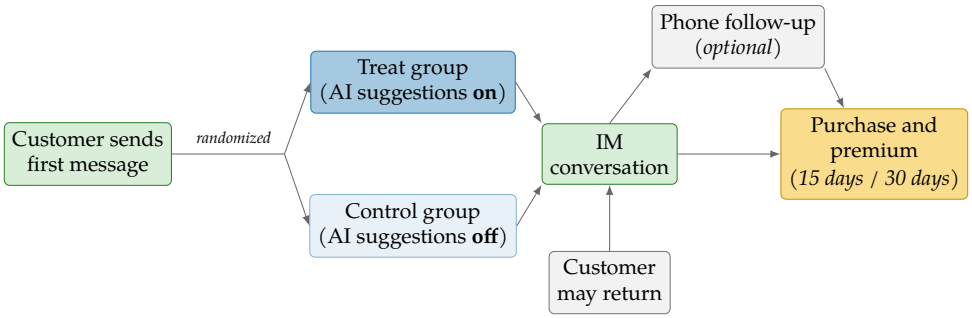


Figure 3: Assignment at the First Message and Outcome Windows

Notes: Customers are randomly matched to brokers. The customer is assigned to the Treat group (AI suggestions on in the broker's workspace) or the Control group (AI suggestions off) at the first message, when the generative-AI tool switches on; about four in five customers are in the Treat group, with a lower share on Day 1. Group and broker are taken from the customer's first chat record. Outcomes run from the first-chat date.

2.4 Data and Samples

The analysis draws on linked datasets that Ant Insurance placed in the Ant Open Research Laboratory sandbox, covering the pilot, the experiment and the first seven weeks of full rollout. Customer-broker-day records give dialogue rounds, same-day purchases and premium by product category, the platform's 15-day purchase and premium fields, and, for the experiment, the group label, customer rating, age and

the AI suggestions the broker received and adopted. A broker-day panel gives service time, customers opened, conversions and premium; a broker roster gives rank, joining date and the pilot indicator. Appendix A describes each dataset and Appendix A the construction of outcomes.

Experiment sample. The customer-level analysis covers customers with a positive count of dialogue rounds on at least one day of the experiment (Figure 2(a)). Three definitions fix the sample:

1. *Group.* Treat equals one if the label on the customer's first chat record is Treat.
2. *First-chat date.* The day of the customer's first chat, including the transition days.[6]
3. *Broker.* The broker on the customer's first chat record.

The experiment sample contains 164,732 customers served by 1,070 brokers. Pilot brokers serve 24.5 percent of the customers in the experiment. For them, a Control assignment withdrew a tool they had used for ten weeks, so their Control customers received a withdrawal rather than the standard service. The main sample is therefore the customers of non-pilot brokers: 124,347 customers served by 831 brokers, of whom 80.4 percent are in the Treat group; 800 of these brokers serve both groups, and the comparison within broker comes from them. Results for all brokers are reported as robustness. Appendix Table A2 summarizes the samples and non-pilot brokers' output in each phase, and Appendix Table A5 gives the steps from the raw records. Output per broker rises from the pilot to the experiment and full rollout, which mixes the rollout with seasonal demand, so I never compare periods directly.[7]

This study was remotely conducted on the *Ant Open Research Laboratory*[8] in an Ant Group Environment. All data was sampled and desensitized, and was analyzed on the *Ant Open Research Laboratory*. The laboratory is a sandbox environment where the author can only remotely conduct empirical analysis and individual observations are invisible.

2.5 Variables

Platform customer rating. Ant Insurance rates each customer B (lowest expected value) < A < S < S+ (highest), with a separate New customer category; I take the rating

[6]For 6,025 customers, the first chat falls on one of the two transition days. Their first-chat date is retained as the starting point for outcomes and date fixed effects.

[7]The product mix shifts with the calendar as well: the life share of policies is 54 percent during the pilot, 68 percent during the experiment and 77 percent in the first three weeks of full rollout among non-pilot brokers, and 55, 68 and 77 percent among pilot brokers, who had the tool in all three periods.

[8]https://www.deor.org.cn/labstore/laboratory. The laboratory follows a "usable but not visible" design: researchers run their own code on the data inside the environment but cannot view, copy or download individual records, and only aggregate results that pass review leave it. The code for every result in this paper is kept in the laboratory and can be rerun there to reproduce the results.

from the first chat record. Among Control customers of non-pilot brokers, B and A account for 86.6 percent. The rating does not respond to assignment (Appendix A).

Customer history classes. I class each customer by the deepest funnel stage reached with any broker in the 70 days before the first chat, a window that starts no earlier than the first day of the pilot. The classes are mutually exclusive:

- **New customers**: no recorded contact in the ten weeks before the first chat.
- **Prior visitors**: customers who had opened a chat window with a broker without sending a message.
- **Prior chatters** (who had not bought): customers who had chatted with a broker but had not bought.
- **Prior buyers**: customers who had bought at least once.

Prior visitors and prior chatters together form the **returning non-buyers**. In the main sample, 57.0 percent of customers are new customers, 22.5 percent prior visitors, 19.0 percent prior chatters and 1.6 percent prior buyers, so returning non-buyers make up 41.5 percent. About a third of customers had earlier contact with the same broker as at the first chat.

Outcomes. All outcomes are measured from the first-chat date (day 0). The main outcomes are *purchase*, an indicator for any purchase on days 0 to 14 (in percentage points), and *premium*, the premium paid over the same days (in yuan). The platform computes these fields from its transaction records over the full 15 days. Most buyers purchase within a few days of the first conversation. The main tables also report a 30-day outcome, which adds the purchases recorded on days 15 to 29, one record per customer and day. I also use purchase and premium by product category and policy size, purchases by day after the first chat, and conversation outcomes. Pilot outcomes are broker averages per calendar day, with inactive days counted as zero. Appendix Table A3 defines every variable.

Broker ability. The main ability index combines non-pilot brokers' performance during the pilot, when they did not have the tool, with tenure at the start of the pilot. It is the first principal component of tenure, mean daily premium, mean daily conversions, and conversions per customer opened; it explains 68 percent of the variance and is standardized across the 820 non-pilot brokers with the necessary inputs and at least ten active days. No experiment-period platform score enters it.[9] Tenure alone is the robustness measure. To compare measures on a common sample, the ability regres-

[9] An earlier comparison index also includes the platform score recorded in the experiment panel, which is not a clean pre-assignment measurement; it appears only as an alternative specification and in the accounting benchmark of Section 7.

sions contain 120,586 customers of 746 brokers, including 50,242 returning non-buyers. The index summarizes existing capability; it does not directly measure the quality of judgment or suggestion screening.

Descriptive statistics. Table 1 describes the main sample. Among Control customers, 3.3 percent buy within 15 days and mean premium is 315 yuan. Premium among buyers is highly right-skewed: among Control buyers the median 15-day premium is 700 yuan, the 90th percentile 20,000 yuan and the mean about 9,600 yuan. I therefore winsorize premium at the 99th percentile of positive values throughout and use Poisson pseudo-maximum-likelihood (PPML) for percentage effects. Appendix Table A4 reports the distribution of every variable.

2.6 Balance Check

Random assignment produces well-balanced groups (Table 1). Customer histories, prior contact with the same broker, visit and chat days, recency, customer ratings, and the broker's tenure and rank are similar across groups, and every standardized difference is below 0.04. Conditional on broker and first-chat-date fixed effects, Treat customers' last earlier contact is 0.5 days more recent, a statistically significant but small difference.

Table 1: Sample Characteristics and Balance

	Control mean (N = 24,344) (1)	Treat mean (N = 100,003) (2)	Difference (3)	p (4)	Norm. diff. (5)
Panel A. *Customer history in the 70 days before the first chat*					
New customers (%)	56.01	57.18	0.33 (0.35)	0.345	0.024
Prior visitors (%)	22.25	22.53	−0.10 (0.30)	0.744	0.007
Prior chatters (%)	20.11	18.74	−0.29 (0.27)	0.270	−0.035
Prior buyers (%)	1.64	1.55	0.06 (0.09)	0.518	−0.007
Prior contact with the same broker (%)	35.44	34.60	0.02 (0.35)	0.958	−0.018
Visit days	1.84	1.76	−0.006 (0.025)	0.807	−0.022
Chat days	0.47	0.43	−0.003 (0.009)	0.747	−0.033
Days since last contact	15.13	14.89	−0.49 (0.20)	0.015	−0.013
Customer rating S+ (%)	5.32	5.05	−0.25 (0.17)	0.135	−0.012
Customer rating S (%)	4.90	4.80	−0.06 (0.16)	0.722	−0.004
Customer rating B (%)	36.93	37.29	0.47 (0.35)	0.191	0.007
New customer rating category (%)	3.10	2.72	−0.17 (0.11)	0.115	−0.023
Premium bought before the first chat (yuan)	435	331	−62.9 (64.9)	0.333	−0.012
Panel B. *Characteristics of the customer's broker*					
Tenure (SD)	0.01	0.00	−0.0040 (0.0077)	0.604	−0.008
Rank L2 (%)	57.00	57.12	0.14 (0.41)	0.727	0.002
Rank L3 (%)	22.43	22.37	−0.04 (0.34)	0.896	−0.001
Rank L4 (%)	9.83	9.78	−0.01 (0.27)	0.979	−0.002
Rank L5 (%)	1.98	2.06	0.10 (0.10)	0.347	0.006
Panel C. *Outcomes, for scale*					
Purchase on the first-chat day (%)	1.96	2.05	0.16 (0.10)	0.091	
Purchase, 15 days (%)	3.28	3.30	0.12 (0.13)	0.354	
Premium, 15 days (yuan)	315.40	377.80	88.4 (40.2)	0.028	
Chat rounds	5.50	5.45			
Chat days	1.36	1.35			

Notes: Customers of non-pilot brokers who chatted at least once in the experiment. 831 brokers, 800 of whom serve both groups. History classes use contact with any broker in the 70 days before the first chat. Column (3): coefficient on Treat with broker, first-chat-day and customer-rating fixed effects (Panel A; rating rows without customer-rating fixed effects) or first-chat-day and customer-rating fixed effects (Panel B); standard errors clustered by broker in parentheses; percentages in percentage points. Column (5): difference in means over the square root of the average group variance. Days since last contact: returning customers only. Prior premium: pilot period. Tenure is standardized across brokers. Panel C: column (3) is the Treat coefficient with broker, first-chat-day and customer-rating fixed effects, as in Table 2. No stars.

3 Access to AI Advice and Insurance Purchases

This section examines how access to AI advice changes whether customers buy insurance, how much they pay and which products they buy. Access raises premium per customer by about 28 percent and the probability of buying a life policy by 14 percent. The premium effect comes from more customers paying 10,000 to 50,000 yuan, and the extra life buyers come mostly from customers who would not have bought at all. The effect arises on the day of the first chat, through conversion within the conversation.

3.1 Empirical Strategy

Because assignment was random within broker and first-chat date, comparing Treat and Control customers of the same broker who first chatted on the same day identifies the intent-to-treat effect of access to AI advice. For customer i, served by broker $b(i)$, with first-chat date $t(i)$ and customer rating $g(i)$, I estimate

$$y_i = \beta \, \text{Treat}_i + \alpha_{b(i)} + \delta_{t(i)} + \mu_{g(i)} + \varepsilon_i, \tag{1}$$

where y_i is an outcome measured from the first-chat date and Treat_i is the group on the first chat record. Broker fixed effects make the comparison one between customers of the same broker, first-chat-date fixed effects absorb common calendar-day conditions, and customer-rating fixed effects improve precision. Standard errors are clustered by broker. The main sample is the 124,347 customers of the 831 non-pilot brokers who chatted at least once in the experiment; the 164,732 customers of all 1,070 brokers form a robustness sample. The coefficient β is the effect of deployment as it occurred, including conversations that ended before a suggestion was produced.[10] For premium I also estimate a Poisson pseudo-maximum-likelihood model with first-chat-date and customer-rating fixed effects and report $100(e^{\beta} - 1)$; Section 4 adds broker fixed effects to this model.

Outcomes. The primary outcomes are the platform's 15-day fields: whether the customer bought any policy, and the premium paid. The platform computes them for every customer from its transaction records, and they count only policies that were not refunded. Product categories and policy sizes come from the daily purchase records. Premium is winsorized at the 99th percentile of positive values (200,000 yuan for the 15-day field).

[10] I retain these customers and do not condition on realized suggestion receipt or conversation length, which can respond to assignment.

Inference. Ant Insurance designed and ran the experiment; my analysis was not pre-registered. I therefore designate a primary family of three tests: (M1) the average effect on 15-day premium; (M2) the interaction of Treat with pre-experiment broker ability in 15-day premium; and (M3) the difference between the life and the health purchase interactions with ability (M2 and M3 are in Section 4). The family is estimated on the common sample with the ability index (120,586 customers of 746 brokers), with p-values from a wild cluster bootstrap that imposes the null (Rademacher weights by broker, 999 draws) and a Holm adjustment. All other estimates are secondary. The estimand is realized premium per customer, an additive revenue measure that also enters the decomposition and the model; under random assignment the coefficient in equation (1) estimates it without functional-form assumptions, so the primary tests use it. Because most customers pay nothing and premium among buyers is heavily right-skewed, I also report Poisson pseudo-maximum likelihood (PPML; Santos Silva and Tenreyro, 2006) with the same fixed effects, including broker effects, which measures each effect in proportion to the Control mean. With the Poisson statistics and a score bootstrap that imposes the null, the same family gives Holm-adjusted p-values of 0.066 for all three tests (Appendix Table A9, Panel C). For the average effects I also report randomization inference that permutes Treat within broker × first-chat-date cells, with the coefficient of equation (1) as the statistic (4,999 draws; Appendix Table A8); it tests the sharp null of no effect for any customer. The Romano and Wolf adjustment is in Appendix Table A9.

Assignment. Treat and Control customers have similar histories and broker characteristics before the first chat (Section 2.6). Brokers knew which customers were in the Treat group, and about 80 percent of each broker's customers were. The contrast is therefore the effect of access in one conversation for a broker who also uses the tool in most other conversations. Effort moved from Control to Treat customers would inflate it; skills from the tool carried into Control conversations would attenuate it. Section 3.7 tests both.

3.2 Purchases and Premium

Table 2 reports the main estimates. Premium is the amount paid for retained policies bought in the outcome window: realized premium revenue and household insurance expenditure, not broker commissions or profit.

15 days. Access raises premium per customer by 88 yuan (standard error 40), or 28 percent of the Control mean of 315 yuan (Panel A). The randomization-inference p-value is 0.064, because the permutation distribution of the premium coefficient has heavier tails than a normal. In the primary family the same effect, 93 yuan on the ability-index sample, has a bootstrap p-value of 0.039 and a Holm-adjusted p-value of 0.044;

across the three primary tests the Holm-adjusted p-values are 0.030 to 0.044 with OLS and 0.066 with PPML. The probability of any purchase rises by 0.12 percentage points on a Control rate of 3.28 percent, which is not statistically significant. PPML with broker fixed effects gives the same proportional effect, 28 percent ($p = 0.040$; randomization inference $p = 0.085$).

30 days. Premium rises by 152 yuan (standard error 60), 29 percent of the Control mean of 529 yuan, and the purchase probability by 0.28 percentage points on 3.85 percent (Panel B). Both reject the sharp null under randomization inference ($p = 0.025$ and 0.041). The PPML effect on 30-day premium is 29 percent ($p = 0.016$; randomization inference $p = 0.028$).

Life insurance. The probability of buying a life policy within 15 days rises by 0.33 percentage points (standard error 0.11), 14 percent of the Control rate of 2.36 percent, with a randomization-inference p-value of 0.002; over 30 days it rises by 0.47 points, 17 percent of 2.69 percent. This is the most robust single estimate in the experiment. Health-insurance purchases do not change (Table 3).

Customer history. Returning non-buyers, customers who had visited or chatted with a broker in the ten weeks before the first chat but had not bought, show effects of similar proportional size (Panels C and D). Premium rises by 92 yuan for new customers (standard error 46) and by 113 yuan (57) for returning non-buyers, about 38 percent in both classes; among returning non-buyers, prior chatters show the largest life effect, 0.81 percentage points (0.28). Both classes gain more in percent than the full sample because its Control mean includes prior buyers, 1.6 percent of customers with a Control mean of 3,356 yuan and an imprecise effect of 48 yuan (standard error 1,028); the class estimates come from separate regressions, so they need not average to the pooled 88 yuan. These differences, and those across time since last contact, customer ratings and age bands, are not statistically significant (Appendix B).

All brokers. The sample of all brokers gives similar or somewhat smaller effects (Panels E and F).

Table 2: Effects of Access to AI Advice on Purchases and Premium

Dependent variable:	Premium PPML (% change) (1)	Premium (yuan) (2)	Purchase (pp) (3)	Purchase, first-chat day (pp) (4)	Life premium PPML (% change) (5)	Life premium (yuan) (6)	Life policy (pp) (7)
Panel A. *Non-pilot brokers, 15 days*							
Treat	28.2**	88**	0.12	0.16*	17.1	172	0.33***
	(15.5)	(40)	(0.13)	(0.10)	(14.0)	(116)	(0.11)
Control mean		315	3.28	1.96		958	2.36
Panel B. *Non-pilot brokers, 30 days*							
Treat	29.0**	152**	0.28**		16.6	205	0.47***
	(13.7)	(60)	(0.14)		(12.6)	(130)	(0.12)
Control mean		529	3.85			1,177	2.69
Panel C. *Returning non-buyers, 15 days*							
Treat	39.0*	113**	0.11		36.2*	371*	0.39**
	(24.4)	(57)	(0.20)		(25.4)	(194)	(0.18)
Control mean		297	3.24			993	2.48
Panel D. *Returning non-buyers, 30 days*							
Treat	40.1**	214**	0.40*		29.1	388*	0.58***
	(23.0)	(101)	(0.22)		(21.3)	(229)	(0.19)
Control mean		569	3.96			1,322	2.90
Panel E. *All brokers, 15 days*							
Treat	22.5**	77**	0.02	0.12	17.0	172*	0.32***
	(12.5)	(36)	(0.11)	(0.09)	(12.1)	(100)	(0.09)
Control mean		338	3.37	2.02		963	2.35
Panel F. *All brokers, 30 days*							
Treat	21.4**	126**	0.18		15.8	200*	0.43***
	(11.1)	(55)	(0.12)		(10.6)	(112)	(0.10)
Control mean		582	3.94			1,200	2.69
Broker, first-chat-day, rating FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Customers who chatted at least once in the experiment: Panels A and B, 124,347 customers of 831 non-pilot brokers; Panels C and D, the 51,579 returning non-buyers among them; Panels E and F, 164,732 customers of all 1,070 brokers. OLS with broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. Premium in yuan, winsorized at the 99th percentile of positive values. Columns (1) and (5): Poisson pseudo-maximum likelihood with broker, first-chat-day and customer-rating fixed effects (broker effects by the exact concentrated likelihood; brokers, days or ratings with no premium drop out), reported as $100 \times (\exp(\beta) - 1)$ with delta-method standard errors; stars test the Poisson coefficient. Randomization-inference p-values are in Appendix Table A8. Stars refer to the cluster tests. Primary family (wild cluster bootstrap with the null imposed, Holm adjustment): see Section 3.1; on the common ability-index sample the 15-day premium effect is 93 yuan, bootstrap $p = 0.039$, Holm $p = 0.044$; with the PPML statistics the same family gives Holm $p = 0.066$ for all three tests. Robustness and multiple testing: Appendix Tables A6 and A9. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

3.3 Where in the Premium Distribution

With 97 percent of Control customers paying nothing within 15 days, the average premium effect comes from a small share of customers. Figure 4 plots the effect of access on the probability that premium exceeds a threshold k, for k from 0 to 100,000 yuan. Effects at thresholds up to 5,000 yuan are close to zero. They become positive and significant from 10,000 yuan: at 20,000 yuan the share of customers paying more rises by 0.115 percentage points (standard error 0.041) on a Control rate of 0.32 percent, and at 50,000 yuan by 0.078 points (0.028) on 0.12 percent, increases of 36 and 63 percent. Above 100,000 yuan the effect is small and not significant. Returning non-buyers show the same shape. Integrating the threshold effects gives an implied 15-day effect between 55 and 106 yuan, around the regression estimate. Because the upper tail matters, the mean depends on the cap: caps between the 97.5th and 99.5th percentiles give significant effects of 70 to 102 yuan, while the uncapped mean is not statistically significant (Appendix Table A7). Table 3 gives the complementary view by interval: effects on life-policy purchases within each premium bin (Section 3.4).

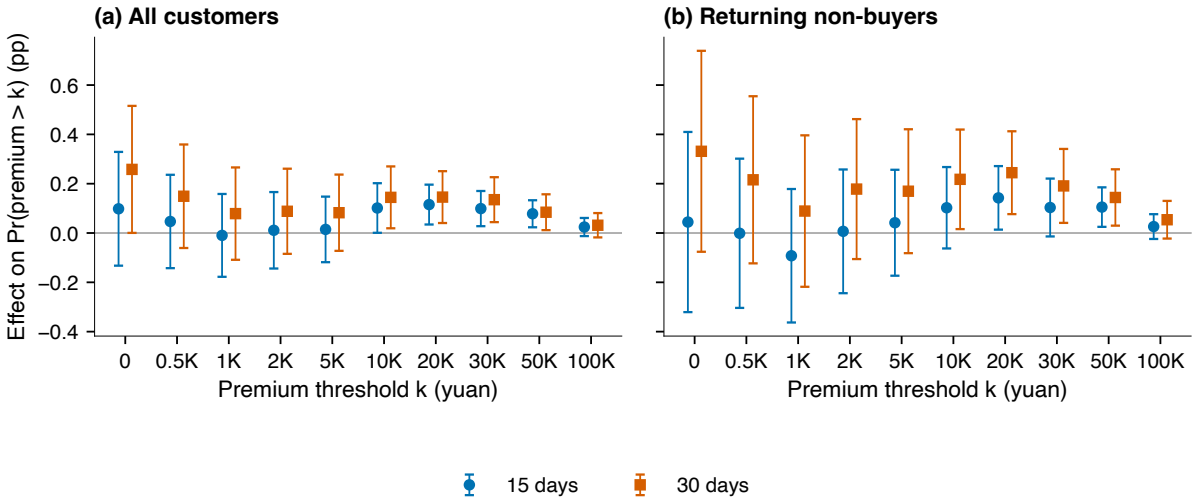


Figure 4: Effects on the Premium Distribution

Notes: Coefficient on Treat, in percentage points, for an indicator that premium within 15 days (circles) or 30 days (squares) exceeds k yuan; K denotes 1,000 yuan. Panel (a): all 124,347 customers of non-pilot brokers; panel (b): the 51,579 returning non-buyers. Joint tests of no effect at all ten thresholds: $p = 0.050$ and 0.125 in panel (a), 0.087 and 0.109 in panel (b), for 15 and 30 days. Broker, first-chat-day and customer-rating fixed effects; 95 percent confidence intervals clustered by broker.

Purchase probability and premium per buyer. Higher premium per customer can come from more buyers or from larger purchases per buyer. For new customers, whose premium rises by 92 yuan (Section 3.2), Control levels from the regression and a broker bootstrap split the effect into 19 yuan from purchase probability (95 percent interval −8 to 48) and 73 yuan from premium per buyer (−12 to 159). Both intervals include zero, so I do not report a precise share; the point estimates assign about four-fifths of

the effect to premium per buyer, as does the same split for returning non-buyers served by the strongest brokers (Section 4). Premium per buyer is not itself a randomized comparison; the threshold effects above are.

3.4 Product Choice and Premium Size

The additional purchases are in life insurance. Table 3 reports effects by product and by life-policy size. Health-insurance estimates are close to zero (-0.026 and -0.003 percentage points at 15 and 30 days). The life increase spans small and middle-sized policies: 0.14 points below 10,000 yuan, 0.09 points at 30,000 to 50,000 yuan and 0.08 points at 50,000 to 100,000 yuan, against Control rates of 1.36, 0.12 and 0.26 percent. The increase below 10,000 yuan does not conflict with the near-zero threshold effects up to 5,000 yuan in Section 3.3: the bins count life policies, the thresholds use each customer's total premium across products, and the small-policy estimate is significant only at the 10 percent level (standard error 0.084). Estimates above 100,000 yuan are small and not statistically significant. These bins refer to premiums paid, not sums insured. The health estimate is also close to zero within every health premium range, including policies of 1,000 yuan or more, which are mostly long-term critical-illness cover (all $p > 0.2$).

Expansion or substitution. The induced life buyers could be customers who would otherwise have bought nothing or customers who would otherwise have bought health insurance. Few customers buy both lines: from the pilot to the end of the rollout window, 4.6 percent of life buyers also bought health insurance and 7.2 percent of health buyers also bought life insurance (Appendix Table A1). Splitting customers into mutually exclusive baskets over days 0 to 14 gives a life effect of $+0.35$ percentage points (standard error 0.11), a no-purchase effect of -0.30 (0.15) and a health-only effect of -0.03 (0.08) (Appendix Table A13). The no-purchase share falls by almost as much as the life share rises, so the data lean toward market expansion, but they do not identify it: random assignment fixes each basket's share, not how customers move between baskets. Only the assumption that access moves customers only toward life yields a point estimate, 83 percent of induced life buyers from no purchase and 10 percent from health only.

A sharper test looks where diversion would have to occur. If access moved customers from health to life insurance, health purchases would fall most among customers who were most likely to buy health insurance. I predict each customer's probability of buying health insurance within 15 days from characteristics fixed before assignment, using Control customers only and cross-fitting, and estimate effects by quintile of this prediction (Appendix Table A14). Health purchases do not fall in any quintile; in the top decile the health estimate is $+0.18$ percentage points (standard error 0.43) on a

Control rate of 3.1 percent. Life purchases rise most in the top quintile and decile, by 0.82 and 1.38 points on Control rates of 3.6 and 4.5 percent. Customers inclined toward health insurance thus buy more life insurance without buying less health insurance, which is consistent with expansion rather than diversion.

Table 3: What Customers Buy: Product Type and Policy Size

	15 days		30 days	
	Treat effect	Control mean	Treat effect	Control mean
Dependent variable	(1)	(2)	(3)	(4)
Panel A. *Purchase by product and life-policy size (pp)*				
Health policy	−0.026	1.31	−0.003	1.47
	(0.083)		(0.089)	
Life policy	0.331***	2.36	0.468***	2.69
	(0.109)		(0.117)	
Under 10k yuan	0.145*	1.36	0.223**	1.54
	(0.084)		(0.090)	
10k to 30k yuan	0.076	0.49	0.080	0.60
	(0.053)		(0.058)	
30k to 50k yuan	0.089***	0.12	0.083***	0.16
	(0.028)		(0.032)	
50k to 100k yuan	0.075**	0.26	0.089**	0.32
	(0.037)		(0.041)	
100k to 200k yuan	0.002	0.21	0.029	0.23
	(0.033)		(0.035)	
200k yuan or more	0.015	0.12	0.026	0.13
	(0.025)		(0.027)	
Panel B. *Buyers only (descriptive)*				
Life share of purchase records (pp)	3.59**	58.5	4.42***	58.9
	(1.83)		(1.70)	
Purchase records per buyer	0.014	1.17	−0.003	1.25
	(0.018)		(0.023)	
Premium per purchase record (yuan)	314	20,700	117	21,100
	(2,222)		(2,073)	

Notes: Customers of the 831 non-pilot brokers who chatted at least once in the experiment (124,347). OLS with broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. Purchase records with a product category, one per customer, day and category. Panel A: life-policy size bins of single-record premium (not exclusive across customer purchases); Control means in percent. Panel B conditions on buying (4,999 and 5,719 buyers), so it is descriptive; premium per policy winsorized at the 99th percentile of positive values. All brokers give the same pattern. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

3.5 When the Purchase Effect Arises

The purchase effect arises on the day of the first chat. Access raises purchases on that day by 0.18 percentage points (standard error 0.10) and does not change the first-purchase

hazard on later days. A daily model of days 0 to 14, which separates keeping customers in the conversation from converting conversations into purchases, attributes about 85 percent of the modelled 15-day purchase effect to the purchase hazards and about 14 percent to continuation (Appendix B).

Access adds buyers; it does not make them decide sooner. Among customers who buy life insurance within 29 days, 39 percent of Treat buyers and 40 percent of Control buyers buy on the first-chat day, and the median buyer in both groups buys on day 1; for savings-type life insurance the median is day 3 in both groups (Appendix Table A15). The life-insurance effect is 0.16 percentage points on day 0 and 0.33 points by day 14, while the health-insurance effect stays near zero at every horizon.

The effect of initial access also accumulates. Figure 5 cumulates the same customers' purchase records from the first-chat day; during the experiment their original assignment governs access in later conversations, and after full rollout brokers have the tool for both groups. The purchase estimates build to 0.43 percentage points after 30 days for all customers and 0.56 points for returning non-buyers.[11] Cumulative purchases remain 0.48 points higher by full-rollout week 7 (Appendix Table A16). This persistent gap after common access limits the explanation that access only moved purchases forward in time.

[11]These cumulative estimates use purchase records only, one per customer and day; the 30-day outcome in Table 2 uses the platform 15-day field, so the two measures differ by definition.

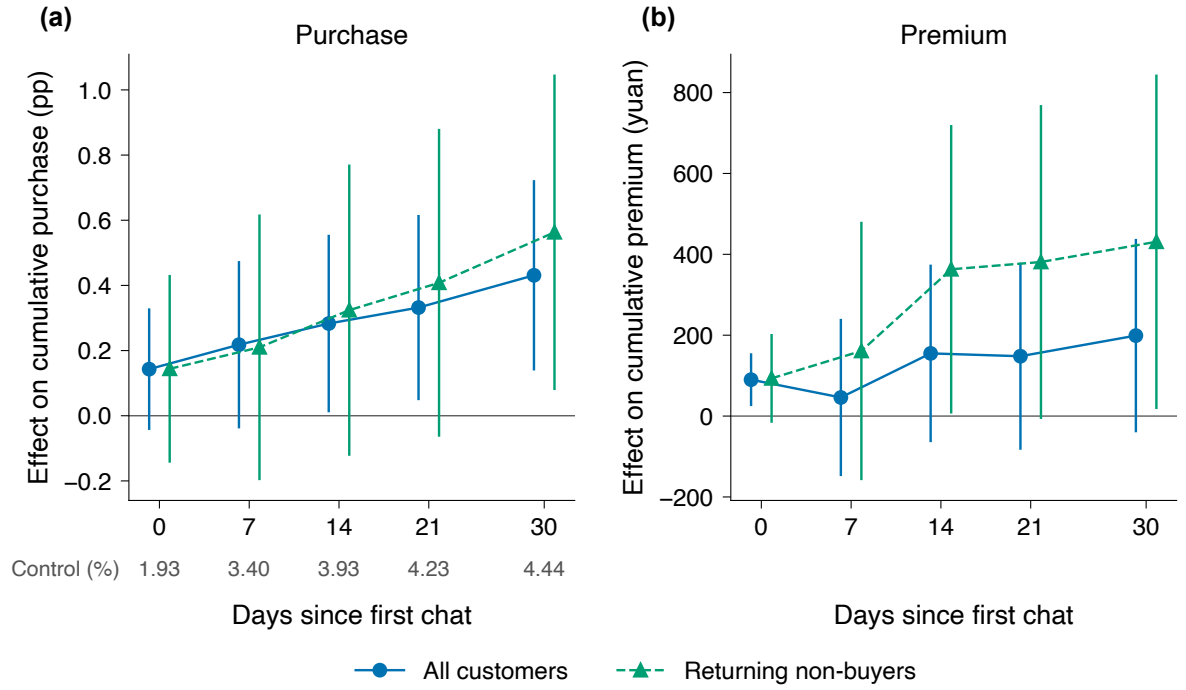


Figure 5: Cumulative Effects of Initial Access to AI Advice

Notes: Customers of non-pilot brokers. Each marker is the coefficient on the customer's initial Treat assignment, with a 95 percent confidence interval, for (a) an indicator of any purchase and (b) premium (yuan, winsorized), cumulated from the first-chat day to each subsequent horizon. Outcomes use purchase records, one per customer and day. Circles and solid line: all customers; triangles and dashed line: returning non-buyers. Numbers below the horizontal axis of panel (a): Control mean of cumulative purchase for all customers (percent). Broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker.

3.6 Subsequent Customer Contact

Treat customers are 1.37 percentage points more likely to return to chat on days 15 to 45, on a Control rate of 15.19 percent (Table 4). The extra buyers stay in contact: the joint probability of buying on days 0 to 14 and returning to chat rises by 0.30 percentage points, and the joint probability of an initial purchase followed by another purchase day by day 45 by 0.13 points. All six contact and joint-event tests survive a Holm adjustment. Joint probabilities do not identify the later behavior of the specific customers induced to buy; further estimates are in Appendix B.

Later purchases are not confined to life insurance. On days 15 to 45, access raises life purchases by 0.17 percentage points (standard error 0.07) on a Control rate of 0.74 percent, and health purchases by 0.08 points (0.03) on a Control rate of 0.18 percent (Appendix Table A14, Panel B). The health product that shows no effect at first thus gains later, once customers have stayed in contact with the broker.

Table 4: Initial Purchases and Subsequent Customer Contact

Dependent variable (pp)	Effect, all	Control	Effect, returning	Control
Purchase, days 0 to 14	0.283**	3.931	0.324	4.151
	(0.139)		(0.228)	
Return chat, days 15 to 45	1.366***	15.190	1.761***	20.600
	(0.266)		(0.443)	
Initial purchase and later chat	0.301***	1.471	0.424***	1.717
	(0.086)		(0.143)	
Initial purchase and another purchase day	0.134**	0.657	0.261***	0.679
	(0.061)		(0.096)	

Notes: All 124,347 customers of non-pilot brokers, including 51,579 returning non-buyers; all customers remain in the denominator. Broker, first-chat-date and customer-rating fixed effects; standard errors clustered by broker in parentheses. Initial purchase is a purchase record on days 0 to 14. Return chat is a positive dialogue-round record on days 15 to 45. Another purchase day occurs after the first purchase and by day 45. Joint outcomes do not condition the estimation sample on buying. The six return-chat, purchase-and-chat and repeat-purchase tests across the two populations all survive Holm adjustment (largest adjusted $p = 0.028$). These outcomes do not identify which customers are induced buyers, satisfaction or suitability.

Consumer welfare. The data cannot measure the welfare value of the additional policies, but they can rule out the most aggressive form of pressure selling. Purchases are retained policies, net of refunds; later contact rises rather than falls; and the gains are not concentrated in the largest policies. New customers do not buy more life policies of 100,000 yuan or more (-0.01 percentage points, standard error 0.05), and prior buyers are less likely, not more, to spend more than twice their own earlier spending (-2.19 points on a base of 4.51 percent, $p = 0.086$). These patterns are not proof of suitability: the records do not measure existing coverage, needs, income or cooling-off cancellations. Commission incentives and the tool's product suggestions should be read

together (Anagol et al., 2017; Egan, 2019; Bergstresser et al., 2009): a tool trained on past successful conversations may learn to propose what sold rather than what fits.

3.7 Spillovers and Robustness

Because brokers knew each customer's group, effort moved toward Treat customers would inflate the estimates and skills carried into Control conversations would attenuate them (Section 3.1). The daily mix of a broker's new chats tests both: within a broker and day, the number of Treat customers varies randomly around the day's Treat share. A day with more Treat customers should therefore leave the broker's Control customers worse off under effort reallocation and better off under skill spillovers. Neither happens: one more Treat customer on the same day changes Control customers' 15-day premium by 6 yuan (standard error 23), and the implied effect of full access is imprecise but close to the customer-level effect (Appendix B).

The 15-day premium effect is largest for customers who first chatted in week 1 and smaller later, though equality across weeks is not rejected ($p = 0.27$); the decline comes from Treat outcomes falling, not Control outcomes rising (Appendix Figure A10). Adding prior-funnel-stage fixed effects or dropping transition-day customers leaves the premium effect at 88 and 85 yuan (Appendix Table A6). Section 6.5 discusses learning carried across the rollout at the broker level.

4 Broker Ability and the Revenue Effect

Does generative AI spread expertise to weaker brokers, or raise the return to expertise they already have? In customer support, professional writing and software development, AI helps the least experienced workers most (Brynjolfsson et al., 2025; Noy and Zhang, 2023; Cui et al., 2026); among small-business owners it helps the strongest (Otis et al., 2026). In insurance advice the answer depends on the product and on whether the customer already knows the broker. Among customers who had already chatted with the same broker, higher-ability brokers gain more, mainly through large life policies; for new customers the gain varies little with ability, and in health insurance the tool narrows gaps across brokers. Because brokers know which customers have the tool, part of the premium gradient reflects how they divide effort between the two groups (Section 4.4).

I augment equation (1) with assignment interacted with standardized ability:

$$y_i = \beta\,\text{Treat}_i + \theta\,\text{Treat}_i \times \text{Ability}_{b(i)} + \alpha_{b(i)} + \delta_{t(i)} + \mu_{g(i)} + \varepsilon_i. \tag{2}$$

Broker fixed effects absorb the main effect of ability, so θ compares Treat and Control

customers of the same broker. Ability is the index of Section 2.5, built from the ten pilot weeks without any experiment-period score; it combines each broker's sales performance in those weeks (premium, conversions and conversion rate) with tenure, so it measures recent sales performance more than experience. The sample contains 120,586 customers of 746 non-pilot brokers, including 50,242 returning non-buyers. Its average premium effects, 93 yuan at 15 days and 165 yuan at 30 days, are close to those of the full sample of 831 brokers, so restricting to brokers with the index does not select larger effects.

4.1 Levels and Proportions

A tool that raised every broker's revenue by the same percentage would give a positive interaction in levels wherever strong brokers sell more without it. Figure 6 therefore plots Control and Treat premium per customer by ability quintile on a log scale, where a widening gap means a rising Treat-to-Control ratio.

Without the tool, premium per customer rises little with ability: the Control slope is 34 yuan per standard deviation (standard error 30) for all customers and −6 yuan (39) for returning non-buyers. With the tool, it is 123 yuan (22) and 136 yuan (36). The proportional (Poisson, PPML) Treat effect rises from 18 percent in the bottom quintile to 75 percent in the top, and reaches 164 percent in the top quintile of returning non-buyers (Appendix Table A22). At 30 days the Control slope is positive (99 yuan), but the Treat slope is more than twice as steep (235 yuan). The purchase margin differs: stronger brokers convert more customers in both groups (Control and Treat slopes of 0.34 and 0.40 percentage points), with an interaction of 0.03. The tool raises what strong brokers sell, not how often they sell.

4.2 The Ability Gradient

A standard deviation of ability raises the 15-day premium effect by 107 yuan (standard error 42) and the 30-day effect by 155 yuan (61) (Table 5). This is test M2 of the primary family: wild-cluster-bootstrap $p = 0.010$, Holm-adjusted $p = 0.030$. Among returning non-buyers the interactions are 150 yuan (56) and 211 yuan (96), and the purchase interactions 0.37 percentage points at 15 days ($p = 0.052$) and 0.43 at 30 days ($p = 0.048$). In proportions (Table 5, columns 6 and 7), the Treat-to-Control premium ratio rises by 24 percent per standard deviation for all customers ($p = 0.051$; score bootstrap $p = 0.041$, Holm-adjusted 0.066 in the PPML family) and 40 percent for returning non-buyers ($p = 0.035$), and by 18 and 23 percent at 30 days. The gradient is therefore clearer in yuan than in proportions: stronger brokers gain more premium per customer, and the ratio of Treat to Control premium rises with ability by about a quarter per standard

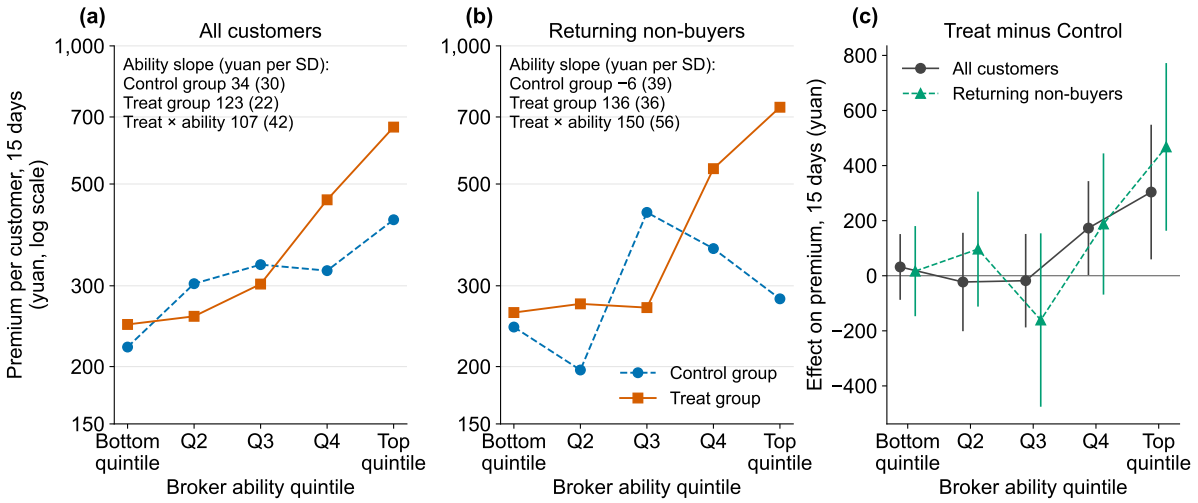


Figure 6: Premium per Customer by Broker Ability, Control and Treat Groups

Notes: Customers of non-pilot brokers with the ability index (120,599 customers of 750 brokers). This sample needs only the index; the common sample of the ability regressions (120,586 customers of 746 brokers) also requires the inputs of every alternative measure. Mean 15-day premium (yuan, winsorized at the 99th percentile of positive values, log scale) by customer-weighted quintile of the pre-experiment ability index (constructed from the pilot weeks). Panel (a): all customers; panel (b): returning non-buyers. Panel (c): the Treat coefficient on 15-day premium within each quintile (broker, first-chat-day and customer-rating fixed effects) with 95 percent confidence intervals clustered by broker (Appendix Table A22). In proportions (Poisson pseudo-maximum likelihood with the same fixed effects), the effects by quintile are 18, −8, −4, 54 and 75 percent for all customers and reach 164 percent in the top quintile of returning non-buyers. Slopes are coefficients on the index estimated within each group with first-chat-date and customer-rating fixed effects; standard errors clustered by broker. The Treat × ability coefficient is estimated within broker (Table 5).

deviation, less precisely.

Tenure alone gives 15-day interactions of 95 yuan overall and 136 yuan among returning non-buyers. The gradient survives dropping one broker at a time (91 to 126 yuan), reweighting to a customer mix common to all quintiles (83 yuan) and removing winsorization (153 yuan) (Appendix Table A23). It reflects general sales ability rather than specialization in life insurance: tenure, conversions and the conversion rate each carry a gradient on their own, specialization carries none, and instrumenting one half of the pilot index with the other raises the gradient to 125 yuan, so measurement error attenuates the main estimate (Appendix Table A20).

The newest brokers gain least. The 15-day premium effect is 23 yuan (51) for brokers with less than three months of tenure and 236 yuan (98) for brokers with more than a year; the linear trend across four tenure groups has $p = 0.090$, and equality across groups is not rejected. These are differences across brokers, not the effect of tenure on one broker, but they run opposite to customer support, where AI lets new workers perform like experienced ones (Brynjolfsson et al., 2025).

The gradient does not come from stronger brokers drawing longer conversations from Treat customers: the Treat interaction on chat-window contact in the first two weeks is -0.62 percentage points, and the premium gradient stays at 108 yuan with realized engagement controls. Nor does it fade: it is 129, 57 and 126 yuan for customers whose first chat fell in weeks 1, 2 and 3 (Appendix Table A31). Where the gradient sits among returning non-buyers, and how much of it reflects the division of effort between groups, is the subject of Section 4.4.

4.3 Where the Gains Go: Products, Policy Size and Task Intensity

Among returning non-buyers, the life-insurance purchase effect rises by 0.49 percentage points per standard deviation of ability, and the life gradient exceeds the health gradient by 0.66 points (wild-cluster-bootstrap $p = 0.002$; Appendix Table A25). For all customers the difference is 0.33 points, test M3 of the primary family (Holm-adjusted $p = 0.044$), and it comes from a negative health gradient (-0.20 points, standard error 0.09) more than a positive life gradient (0.13, 0.11). Figure 7 shows the implied effects at the quartiles and median of ability: life-insurance effects rise with ability, health-insurance effects are small, decline with ability and turn negative for strong brokers. By policy size, the gradient in buying at least 30K yuan is 0.13 percentage points (0.04) for all customers, while the gradient below 5K yuan is -0.09 and not significant.

Skill intensity. To compare tasks directly, I measure for each task cell (life and health purchase, three policy-size bins, any purchase and premium, for new customers and returning non-buyers) how much ability matters without the tool, using non-pilot

Table 5: Pre-Experiment Broker Ability and the Effects of Access to AI Advice

Dependent variable:	Purchase 15 days (pp)	Life purchase 15 days (pp)	Premium 15 days (yuan)	Purchase 30 days (pp)	Premium 30 days (yuan)	Premium 15 days (%) PPML	Premium 30 days (%) PPML
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. *All customers*							
Treat × ability index	0.03	0.13	107**	0.11	155**	23.7*	18.0*
	(0.12)	(0.11)	(42)	(0.14)	(61)	(13.5)	(10.9)
Treat × tenure	0.04	0.04	95**	0.10	91	24.8*	9.7
	(0.13)	(0.12)	(39)	(0.15)	(62)	(15.2)	(11.0)
Control-group ability slope	0.34***	0.23**	34		99**	10.4	18.0**
	(0.11)	(0.10)	(30)		(44)	(9.1)	(8.0)
Treat-group ability slope	0.40***	0.35***	123***		235***	30.3***	34.5***
	(0.06)	(0.05)	(22)		(38)	(5.2)	(5.4)
Panel B. *Returning non-buyers*							
Treat × ability index	0.37*	0.49***	150***	0.43**	211**	40.0**	23.5
	(0.19)	(0.17)	(56)	(0.22)	(96)	(22.3)	(17.6)
Treat × tenure	0.46**	0.25	136***	0.50**	144	43.8*	11.8
	(0.20)	(0.20)	(52)	(0.23)	(104)	(26.7)	(16.9)
Control-group ability slope	0.07	0.04	−6		95	−1.6	15.8
	(0.18)	(0.16)	(39)		(76)	(12.5)	(12.4)
Treat-group ability slope	0.48***	0.52***	136***			31.4***	35.8***
	(0.09)	(0.09)	(36)			(7.9)	(7.4)

Notes: Interaction rows: coefficients on Treat × the broker measure per broker standard deviation, one measure per regression, with broker, first-chat-date and customer-rating fixed effects; broker fixed effects absorb the main effect of ability. Common sample: 120,586 customers of 746 non-pilot brokers, including 50,242 returning non-buyers. Life purchase in column (2) is from the stacked life and health purchase equations of Table A25. Slope rows: the coefficient on the ability index estimated separately within the Control group and within the Treat group, with first-chat-date and customer-rating fixed effects and no broker fixed effects (the index is a broker variable), on the 120,599 customers of 750 brokers with the index. They show how much ability matters without and with the tool; their difference is not the interaction, which is estimated within broker. Blank cells were not estimated or not exported. The index is the first principal component of tenure, mean daily premium, conversions and conversion rate over the ten pilot weeks, built on non-pilot brokers and standardized; no experiment-period platform score enters it. Premium in yuan, winsorized at the 99th percentile of positive values. Columns (6) and (7): Poisson pseudo-maximum likelihood on the same premium outcomes, reported as $100 \times (\exp(\beta) - 1)$ per broker standard deviation with delta-method standard errors; interaction rows with broker, first-chat-date and customer-rating fixed effects on the customers with the index (brokers, days or ratings with no premium drop out), slope rows without broker fixed effects as above; stars test the Poisson coefficient. They measure the gradient in proportions: whether the Treat-to-Control premium ratio rises with ability. With the null imposed, the score bootstrap gives $p = 0.041$ for the all-customer interaction in column (6) and $p = 0.026$ for returning non-buyers. Standard errors clustered by broker in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

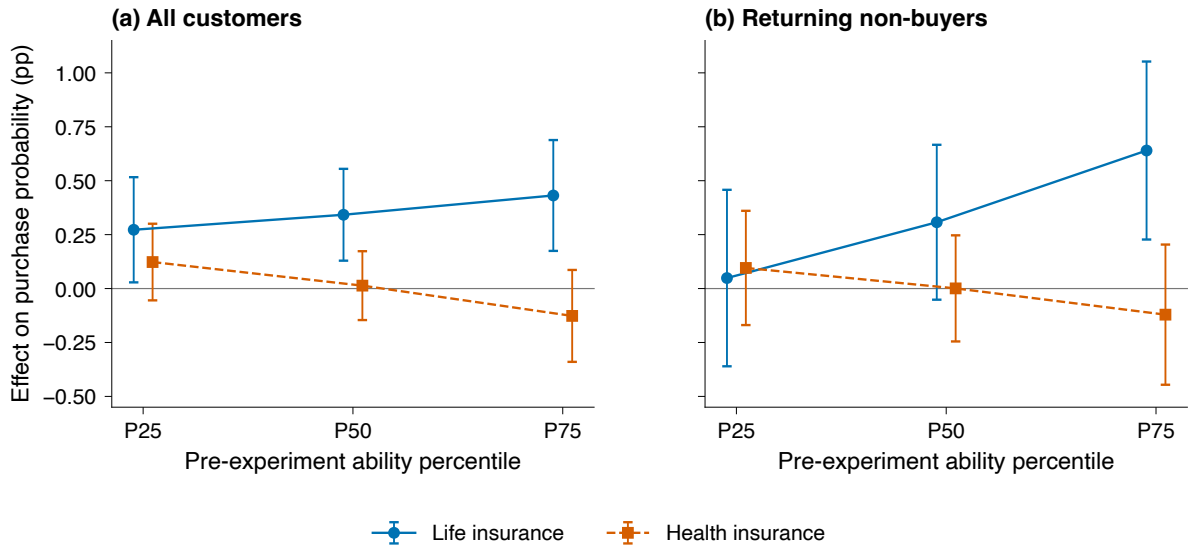


Figure 7: Life- and Health-Insurance Effects across Pre-Experiment Ability

Notes: Stacked product OLS with product-specific broker, first-chat-date and customer-rating effects. Linear combinations of Treat and Treat × the pre-experiment index at customer-weighted P25, P50 and P75; percentage points with 95 percent broker-clustered intervals. Gradient differences are estimated jointly (Appendix Table A25).

brokers' customers whose first chat fell in pilot weeks 6 to 10 and an index rebuilt on pilot weeks 1 to 5. Figure 8 plots the experiment's Treat × ability coefficient against this slope. Ability matters without the tool in all fourteen cells, but the tool does not scale these slopes up uniformly: its interaction is positive in premium, life insurance and policies of 5K yuan and above, and negative in health insurance and new customers' small policies. Among returning non-buyers, cells where ability mattered more without the tool are those where the tool raised its return more (correlation 0.80 across seven cells); for new customers there is no such relation (0.18).

These comparisons are descriptive: cells overlap, few cell interactions are individually significant, and the sign of the cross-cell relation depends on scale (Appendix C). The robust pattern is by the value and complexity of the task: the tool raises the return to ability in life insurance, large policies and among returning non-buyers, and flattens it in health insurance and small policies. The two lines differ mainly in value and in how long customers take to decide: a life policy's median premium is about 20 times a health policy's, and the median life policy is bought 5 days after the first chat against 1 day for health, although the conversations before a purchase are about as long (Appendix Table A1). One explanation is that the tool's content is enough where customers can check the terms themselves, as in short-term medical cover, but in savings-type life insurance customers act on content only as far as they trust the broker who sends it. Section 7 develops this explanation. A second, complementary explanation concerns incentives.

Incentives. The cells that differ in complexity also differ in what a sale pays the broker. Brokers' pay includes a performance component tied to sales, and an average life policy pays the broker more per sale than a health policy because it is much larger: life insurance is 58 percent of the brokers' policies but 91 percent of their premium (Appendix Table A1). Per yuan of premium the ranking reverses, since single-premium savings policies carry first-year commission of only a few percent while long-term protection carries much more. Incentives therefore matter for where the tool raises the return to ability: a broker gains most from using AI suggestions well where a sale is worth most to her, and the decomposition in Section 7.1 shows that strong brokers' gains come mainly from larger purchases per buyer and more life buyers. Incentives cannot produce the gradient through the commission schedule alone, because a broker's Control customers carry the same schedule; but because brokers know which customers have the tool, incentives can shift effort between the two groups, which Section 4.4 examines. Strong brokers also adopt a smaller share of the suggestions they receive (Section 5), which fits selective use. The data contain no commissions, so I cannot separate the tool's accuracy by task from the broker's reward by task; Section 3.6 discusses what the records say about suitability.

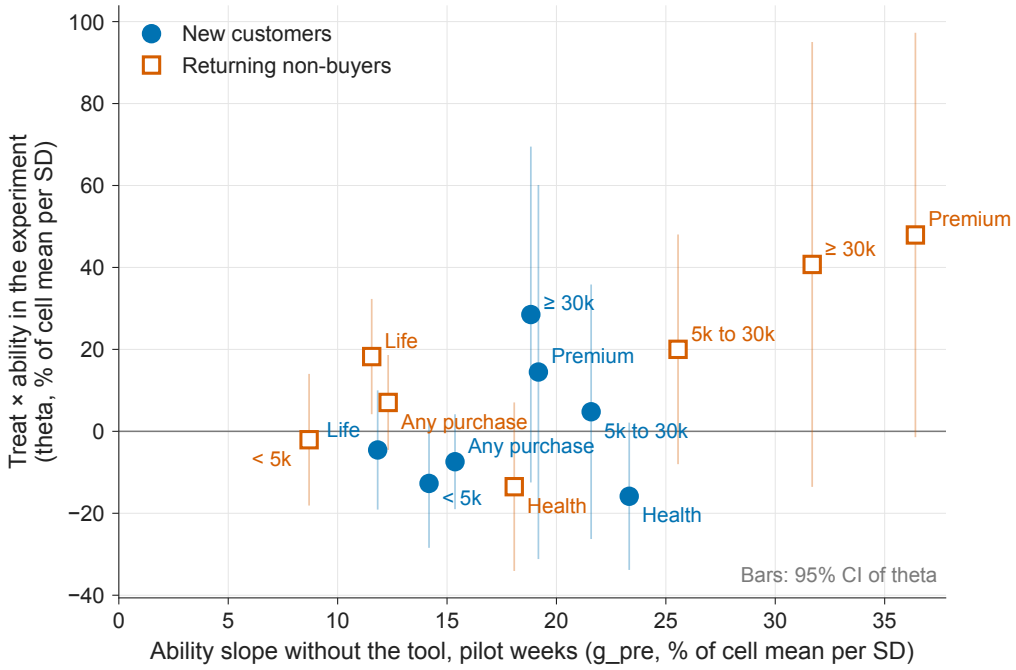


Figure 8: Skill Intensity without the Tool and the Tool's Effect on the Return to Ability
Notes: Each point is a task cell for new customers (circles) or returning non-buyers (squares). Horizontal axis: slope of the 15-day outcome on ability among non-pilot brokers' customers whose first chat fell in pilot weeks 6 to 10, before these brokers had the tool, in percent of the cell mean per standard deviation. Vertical axis: Treat × ability in the experiment, in percent of the Control cell mean, with 95 percent broker-clustered intervals. Ability is the index rebuilt from pilot weeks 1 to 5. K = 1,000 yuan. Estimates in Appendix Table A27.

4.4 Customers Who Know the Broker, and the Division of Effort

The ability gradient among returning non-buyers sits with customers who had already chatted with the same broker, about a third of them. With the main index, the premium gradient is 189 yuan (64) for these customers and an insignificant −59 yuan for the other returning non-buyers ($p = 0.11$ for the difference; Appendix C). In the rest of this section ability is the index rebuilt from pilot weeks 1 to 5 with the main recipe, as in the task cells above. Among these customers, access raises the probability of buying a life policy by 1.04 percentage points (standard error 0.35) on a Control rate of 3.2 percent, and of buying one of 30,000 yuan or more by 0.61 points (0.21) on 1.1 percent; each standard deviation of ability raises these effects by 0.88 (0.33) and 0.56 (0.20) points. Among customers who had chatted only with other brokers, the effects and their gradients are small, and among prior visitors the gradient is smaller; equality across the three groups has p-values of 0.06 to 0.08 for life insurance. The number of earlier chat days with the broker does not change the effect. For new customers, the effect on life purchases does not rise with ability at any policy size (Appendix F).

Brokers see which customers have the tool, so the gradient could partly reflect how they divide effort between the two groups rather than what the tool adds. Two comparisons address this. First, relative to customers of the same brokers who had chatted with them before and whose first chat fell in pilot weeks 6 to 10, before these brokers had the tool, the ability slope in buying a life policy of 30,000 yuan or more falls for Control customers in the experiment (−0.52 of the period mean per standard deviation, standard error 0.17) and is unchanged for Treat customers (0.00, 0.14), and stronger brokers' Control customers buy more health insurance. Much of the experimental gradient among these customers is thus a decline on the Control side; with a common, shorter history window the Treat-minus-Control difference is smaller and not significant. Second, in the pilot every customer of a pilot broker had the tool, so pilot brokers could not divide effort between groups. Measuring ability before the pilot, by an index of the same recipe and by rank, the gradient of pilot status in large life policies among customers who had chatted with the same broker is 0.28 (0.17) and 0.35 (0.17) of the mean per standard deviation, about as large as the experiment's 0.27 and 0.15 with the same measures, but the premium gradient is close to zero (0.04, 0.18, against 0.59, 0.31). These pilot comparisons are imprecise, and they weaken when the outcome window is kept inside the pilot (Appendix F).

I read the evidence as follows. Stronger brokers do convert more large life policies with the tool among customers who know them, even where they cannot favor one group. The premium gradient measured in the experiment also reflects the division of effort between groups, so I treat it as an upper bound on what the tool adds to stronger brokers' premium at full deployment.

4.5 Interpretation

Across these results, access to AI advice raises the revenue return to broker ability in large life policies sold to customers who already know the broker, and narrows gaps in health insurance. With the same tool, brokers and days, the tool amplifies ability in one kind of advice and levels it in another, reconciling the skill-leveling evidence from customer support, writing and software with the evidence for entrepreneurs (Otis et al., 2026). Without the tool, stronger brokers sell larger life policies rather than more of them: among non-pilot brokers' policies during the pilot, the life share of policies varies little with ability (0.39 percentage points per standard deviation, standard error 0.59), while premium per life policy rises by about 26 percent per standard deviation and premium per health policy by about 12 percent. The design does not identify whether strong brokers' advantage comes from evaluating suggestions, adapting them, or the customer context they bring; Section 5 asks what the usage logs can say.

5 AI Adoption and Its Value

Firms, and much research, measure AI by how much it is used. Do strong brokers gain more because they use the tool more? The logs record, for every Treat conversation, whether the broker received AI suggestions and how she handled each one. Adoption turns out to be a stable broker trait but a poor guide to the broker's gain: brokers who adopt more gain less, and gains line up with ability instead.

5.1 Receipt and Adoption

Assignment enables the tool, but generation takes time, and many short conversations end before a suggestion is produced. Receipt therefore reflects the opportunity to receive a timely suggestion, not the broker's willingness to use AI. Let E_i indicate that the broker received at least one AI suggestion while serving customer i. I estimate

$$E_i = \pi \,\mathrm{Treat}_i + \alpha_{b(i)} + \delta_{t(i)} + \mu_{g(i)} + u_i \tag{3}$$

and divide the Treat effect on outcomes by π. This is a Wald scaling per customer whose broker receives suggestions because the tool is enabled; it is a local average treatment effect only under exclusion, which fails if brokers respond to knowing a customer's group.

Assignment raises the share of customers for whom the broker receives any suggestion by 45 percentage points, from 0.7 percent in the Control group, and the share with at least one adopted suggestion by 19 points (Table 6). Scaled by receipt, the premium effect is 196 yuan (standard error 89), and 301 yuan (152) for returning non-buyers.

Receipt rises with ability, from 41 percentage points in the bottom quintile to 50 in the top, while the effect on any adoption stays between 18 and 20 points (Appendix Table A22).

Table 6: Assignment, Receipt of Suggestions, and Revenue

Dependent variable	All customers (1)	Returning non-buyers (2)
Any suggestion: first stage (pp)	44.97***	37.59***
	(0.41)	(0.43)
Control mean (percent)	0.68	1.03
First-stage F	11,931	7,728
Any adoption: first stage (pp)	19.30***	15.30***
	(0.39)	(0.35)
Exposure-normalized outcome effects		
Purchase (pp)	0.26	0.29
	(0.28)	(0.52)
Premium (yuan)	196**	301**
	(89)	(152)

Notes: Customers of non-pilot brokers. First stages are effects of Treat on whether the broker receives or adopts any suggestion for the customer. Exposure-normalized effects instrument any receipt with assignment; units are per additional customer whose broker receives suggestions, under exclusion and monotonicity. They are not returns per individual suggestion. OLS with broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. Premium in yuan, winsorized at the 99th percentile of positive values. A Poisson control-function version of the premium row is too imprecise to report (two-step bootstrap p above 0.8), so the exposure-normalized premium effect is given in yuan only. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

When suggestions arrive. Use is concentrated in the first conversation. Of all suggestions brokers receive for Treat customers over days 0 to 14, 64 percent arrive on the first-chat day, and 89 percent of customers whose broker receives any suggestion receive the first one that day (Appendix Table A33). Suggestions arrive at roughly one per customer message, 0.7 to 1.0 per chat round, so the chance that a broker receives one rises with the length of the day's conversation, from 54 percent with one round to 92 percent with more than 20. The adoption rate moves the other way, from 30 to 8 percent: in long conversations brokers receive many suggestions and send few of them. On the day a customer buys, the chance of receiving a suggestion is 21 percentage points higher than on the day before, but the adoption rate is the same.

5.2 Brokers Who Adopt More Do Not Gain More

Within Treat conversations, more adoption goes with fewer purchases and less premium (Appendix Figure A16). This comparison mixes the value of adoption with the reasons brokers adopt: brokers may lean on suggestions in hard conversations. To learn whether

brokers who adopt more gain more, I therefore compare experimental effects across brokers. Because every broker serves both groups, each broker has her own unbiased premium effect. I split each broker's customers into two halves by the parity of the first-chat day and relate her premium effect in one half to her adoption in the other. The sampling noise in the halves is independent, so the cross-half covariance estimates the covariance of true gain and true adoption, and scaling by cross-half variances gives a noise-corrected correlation.

Across 654 non-pilot brokers, adoption is a stable trait: the adoption rate correlates 0.80 between halves and the share of suggestions sent 0.92 (Appendix Table A29). The broker's premium gain is far noisier: the variance of true broker gains is not significantly above zero, so the corrected correlations are ratios of imprecise quantities, and their intervals can exceed one in absolute value. They nonetheless point one way (Figure 9). The gain correlates -0.39 with the adoption rate (95 percent bootstrap interval -1.63 to -0.11) and -0.33 with the share sent, is unrelated to receipt (-0.04), and correlates positively with ability (0.42, interval 0.08 to 1.55). By tercile of adoption measured in the other half, the 15-day premium effect is 161 yuan (68), 59 yuan and 15 yuan; the top-minus-bottom difference has $p = 0.11$.

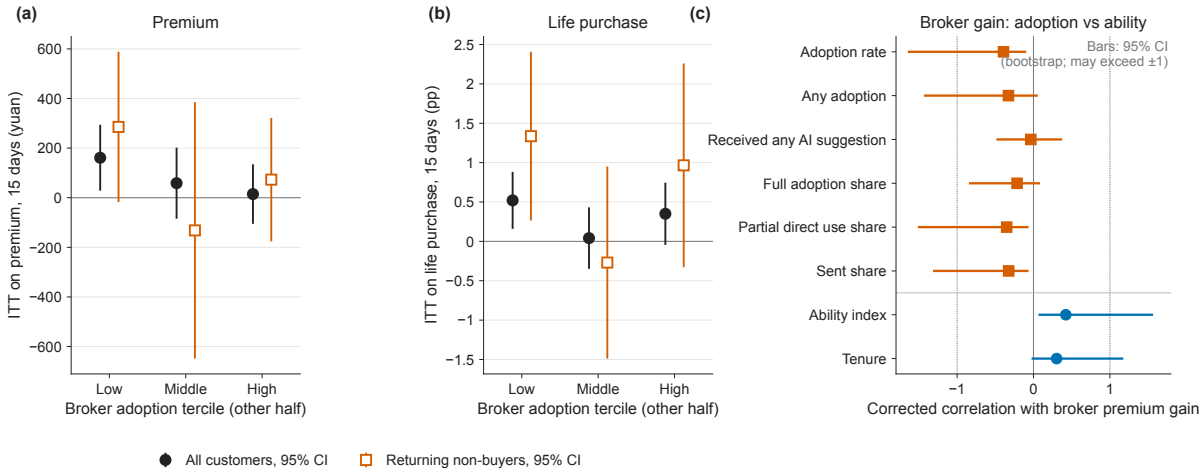


Figure 9: Broker Gains, Adoption and Ability: Split-Sample Evidence

Notes: Customers of non-pilot brokers, split into two halves by the parity of the first-chat day. Panels (a) and (b): Treat effect on 15-day premium (yuan) and life purchase (percentage points) by tercile of the broker's adoption share measured in the other half, for all customers (filled) and returning non-buyers (hollow), with broker-half, first-chat-date and customer-rating fixed effects and 95 percent broker-clustered intervals. Panel (c): corrected correlation between the broker's premium gain in one half and each measure in the other half (adoption measures) or the fixed broker trait (ability index, tenure), for 654 brokers, with 95 percent broker-bootstrap intervals (500 draws), which can exceed ± 1. Estimates in Appendix Table A29.

Broker rank points the same way. Higher-ranked brokers send a smaller share of the suggestions they receive yet have larger premium effects (Figure 10): among returning non-buyers the effect rises by 137 yuan (64) per rank step, although the omnibus test across ranks has $p = 0.089$ (Appendix Table A30). By ability quintile, the share of

received suggestions sent falls from 12.9 to 7.7 percent while the premium effect rises from 32 to 304 yuan. Strong brokers receive more suggestions, send fewer of them and gain the most.

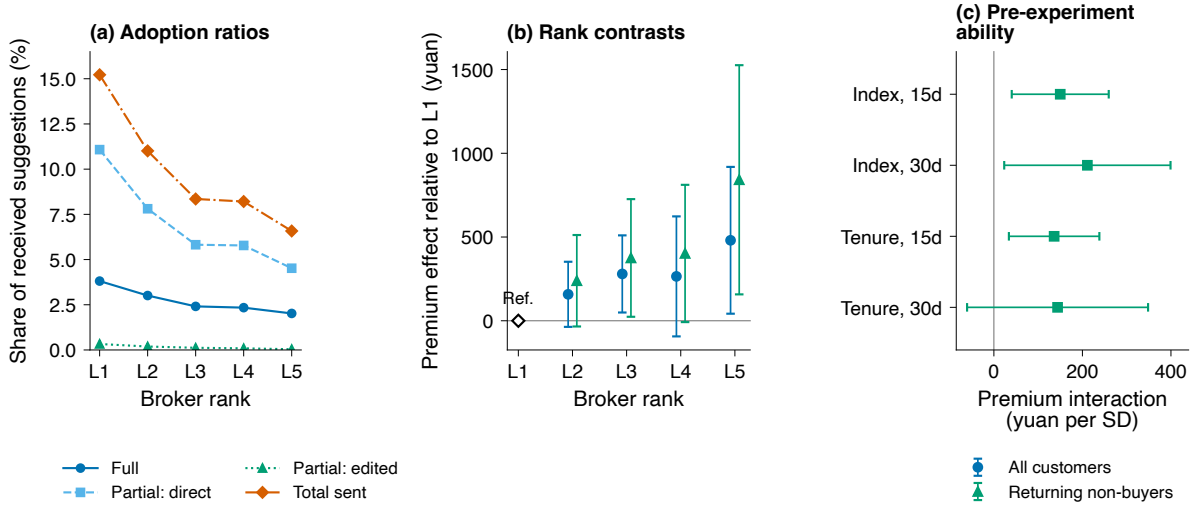


Figure 10: Broker Rank, Adoption, and Pre-Experiment Ability

Notes: Non-pilot brokers. Panel (a): suggestion-weighted full adoption, partial direct sending, partial adoption after edits, and their sum (total sent), divided by receipt; copied-as-inspiration is excluded. Panel (b): 15-day premium-effect contrasts relative to rank L1, estimated jointly; L1 is a reference, not zero effect. Panel (c): returning-non-buyer premium interactions per standard deviation of the pre-experiment index and tenure, at 15 and 30 days. Outcome estimates include broker, date and rating effects; 95 percent broker-clustered intervals. Adoption shares are descriptive.

5.3 Adoption over Time and Broker Time

Over the experiment, adoption and value move apart. Within the same brokers, the share of Treat customers with any adopted suggestion rises from 15.1 percent for week-1 cohorts to 23.4 percent for week-3 cohorts, with unchanged receipt, while the 15-day premium effect is 152, 60 and 10 yuan; equality across weeks is not rejected ($p = 0.27$), so the decline is suggestive (Appendix Table A31). The rise is in the least selective ways of using a suggestion: relative to week 1, the shares of received suggestions that are copied and that are sent directly after deleting parts rise by 5.5 and 2.5 percentage points in week 3, against 0.5 points for full adoption (Appendix Table A34). The increase in copying is broad: 87 percent of brokers copy a larger share in week 3, in every rank. How a broker handled suggestions on the previous day does not predict her gain on the next day's customers.

Nor does the tool show any sign of saving broker time: a random extra Treat customer among a broker's first chats does not lower her chat-window time that day or per customer, and adds a little phone time (Appendix Table A32). Workload instead changes how brokers use the tool. On a day when a broker happens to receive one more

Treat customer than her day's Treat share implies, she adopts 0.18 percentage points more of the suggestions she receives (standard error 0.07, on a mean of 14.9 percent), entirely through suggestions sent directly after deleting parts; full adoption and edited sending do not rise (Appendix Table A34). When busier, brokers rely more on the tool's text and adapt it less, and the gain per Treat customer is smaller on busier days (Section 6.6).

Together, the split-sample, rank, quintile and weekly comparisons show that adoption is a misleading proxy for the value of access to AI advice: strong brokers gain most while adopting least, and adoption rises while the effect does not. One explanation is that strong brokers evaluate and adapt suggestions and send them selectively, as in evidence that the value of discretionary overrides of algorithmic advice varies across decision-makers (Angelova et al., 2026). The usage logs cannot separate this capability from the customer context strong brokers bring, the question Section 4 left open. I return to the gap between adoption and value in Section 7, which sets out a model in which the tool's accuracy differs by product and brokers screen its suggestions.

6 Brokerage Output: Evidence from the Pilot

Sections 3 to 5 compare customers of the same broker over the 20 days of the experiment. That comparison leaves open two questions that the broker-level pilot can answer: whether earlier access changes a broker's premium revenue and the number of customers she serves, and whether any advantage outlasts the period in which only some brokers have the tool. After adjusting for differences that existed before the pilot, pilot brokers do not earn more in the first five pilot weeks; a premium advantage appears from week 6 and persists, at a similar proportional size, after every broker has the tool. Pilot brokers record only slightly more work than other brokers, and most of the advantage is more premium per unit of recorded work, earned mainly through larger purchases. The design cannot tell whether this persistence reflects learning and a larger customer stock or differences in unobserved effort.

6.1 Design, Frame and Balance

Pilot brokers were drawn at random within rank, with probabilities that differ across ranks. Two features of the roster complicate the comparison. First, the pilot changes who works: on the roster at the start of the pilot, pilot brokers are 12 to 20 percentage points more likely to work in each period (Appendix Table A37), so revenue gaps on the full roster mix output with activity, and restricting to brokers active in the first pilot week selects on a pilot outcome. Second, almost all pilot brokers had been active just before the pilot (99.5 percent, against 52.5 percent of other brokers), so the draw appears

to have been made among active brokers.

I therefore define the randomization frame before the pilot: brokers of ranks L2 to L5 who worked on at least three of the nine days before the pilot, 862 brokers of whom 187 are pilot brokers. These nine days are used only to define this frame, to test balance and to form baseline covariates. Pilot shares in this frame are 19.5 percent at L2, 25.2 at L3, 20.7 at L4 and 44.0 at L5 (Table 7, Panel B).

Within rank, pilot brokers were not balanced on pre-pilot effort. They spent more time on the telephone (0.15 log points, standardized difference 0.20) and in chat (0.09 log points, 0.14), and were active on slightly more days, while their premium, conversions, customers opened and tenure do not differ significantly; the seven covariates are jointly imbalanced ($\chi^2 = 29.6$, 7 degrees of freedom, $p = 0.0001$). I adjust for the imbalance by ANCOVA: for broker b and period p,

$$\bar{y}_{bp} = \gamma_p \,\mathrm{Pilot}_b + \alpha_{r(b)} + \phi_p \,\bar{y}_{b0} + X'_{b0}\delta_p + e_{bp}, \tag{4}$$

where $\bar{y}_{bp}$ is the broker's average per calendar day in period p (inactive days count as zero), $\alpha_{r(b)}$ are rank fixed effects, $\bar{y}_{b0}$ is the same outcome over the nine pre-pilot days and X_{b0} are the pre-pilot covariates of Table 7. The periods are the first and second five weeks of the pilot, the experiment, the first three weeks of full rollout and the later weeks. I report robust standard errors and randomization-inference p-values from 500 permutations of pilot status within rank. Adjustment removes the observed imbalance, not differences in unobserved effort correlated with it.

The meaning of the pilot coefficient changes over the rollout. During the pilot only pilot brokers have the tool; during the experiment every broker can use it for Treat customers, and in full rollout for every customer. From the experiment on, γ_p therefore measures the persistence of an early-access advantage, not the effect of current access.

6.2 Revenue by Period

Pilot brokers' premium advantage appears in the second half of the pilot and persists after every broker has the tool (Table 8 and Figure 11). Conditional on pre-pilot sales and effort, pilot brokers earn 536 yuan more premium per day in the first five pilot weeks, a difference that is not statistically significant (randomization-inference $p = 0.29$). The difference rises to 1,076 yuan in weeks 6 to 10 ($p = 0.054$), about 18 percent of the non-pilot mean, and to 3,632 yuan during the experiment ($p = 0.046$), about 24 percent of a non-pilot mean that has nearly tripled with seasonal demand. It remains at 2,924 yuan in the first three weeks of full rollout ($p = 0.13$) and 3,859 yuan in the later weeks ($p = 0.012$), about 14 and 23 percent of the non-pilot means. Conversions and customers opened follow the same pattern with less precision. Without baseline adjustment, and

Table 7: Pilot Assignment and Balance before the Pilot

	Pilot mean (1)	Non-pilot mean (2)	Difference within rank (3)	Standardized difference (4)
Panel A. *Outcomes and effort on the nine days before the pilot*				
Log premium per day	6.50	6.03	0.341	0.12
			(0.239)	
Conversions per day	0.44	0.40	0.022	0.07
			(0.026)	
Customers opened per day	20.6	19.6	0.54	0.09
			(0.48)	
Active-day share	1.000	0.997	0.003**	0.11
			(0.001)	
Log phone seconds per day	7.54	7.38	0.150***	0.20
			–	
Log chat seconds per day	8.10	7.98	0.091**	0.14
			(0.043)	
Tenure (months)	11.5	10.8	−0.33	−0.04
			(0.44)	
Joint test, χ^2 (7 df)				29.6, $p = 0.0001$
Panel B. *Pilot share by rank*				
	L2	L3	L4	L5
Brokers in the frame	544	206	87	25
Pilot brokers	106	52	18	11
Pilot share (%)	19.5	25.2	20.7	44.0

Notes: Pre-pilot frame: brokers of ranks L2 to L5 on the roster at the start of the pilot who worked on at least three of the nine days before it (862 brokers, 187 pilot brokers). These nine days are used only to define this frame, test balance and form the covariates of Table 8. Column (3): pilot minus non-pilot brokers with rank fixed effects, robust standard errors in parentheses (the standard error for phone time was not recovered from the exported output). Column (4): difference divided by the pooled standard deviation. Per-day outcomes count inactive days as zero; phone and chat seconds are recorded broker work time. The joint test covers the seven rows of Panel A. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

in the frame of brokers active in the first pilot week, the gaps are similar; the full roster gives gaps 1.5 to 2 times as large because it also counts the effect of the pilot on whether brokers work (Appendix E).

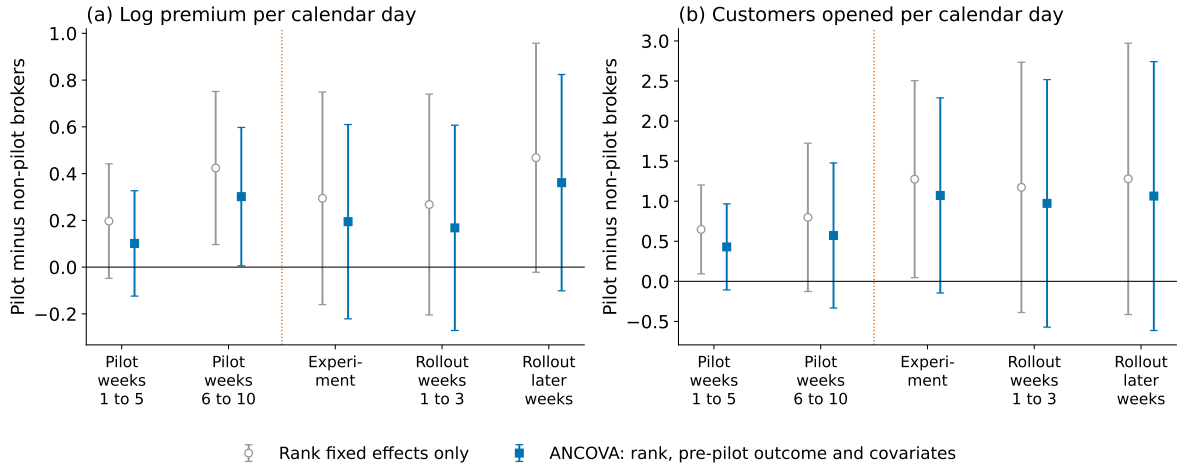


Figure 11: Early Access, Broker Revenue and Customers Served

Notes: Pilot minus non-pilot brokers by period in the pre-pilot frame (862 brokers, 187 pilot brokers), per calendar day with inactive days counted as zero. Hollow circles: rank fixed effects only. Squares: ANCOVA with rank fixed effects, the outcome's pre-pilot value and pre-pilot covariates (Table 8). Bars are 95 percent confidence intervals from robust standard errors. The dotted line separates the pilot, when only pilot brokers had the tool, from the periods in which every broker had it.

6.3 No Convergence after Rollout

If the pilot gap reflected only current access, it would shrink once non-pilot brokers received the tool. It does not: between the experiment and later full rollout the gap changes by 443 yuan (randomization-inference $p = 0.83$) and 0.17 log points ($p = 0.35$), and in logs the gap is no larger in late full rollout than in pilot weeks 6 to 10 (0.04, $p = 0.84$; Appendix Table A38). The proportional advantage that appears in the second half of the pilot is stable through the experiment and full rollout.

Customers served and current-access quality explain part of the gap during the pilot but leave most of it unexplained from the experiment on (Appendix E). Within broker, customers opened per day rise with conversions in the previous seven days, equally for pilot and non-pilot brokers, a feedback channel through which an early gain in conversions can compound into more customers served (Section E.3). Scaling the customer-level effect to a daily flow, current-access quality and extra customers served account for part of the pilot-period gap, while from the experiment on most of the gap is a residual that neither explains (Section E.4).

Table 8: Early Access and Broker Revenue in the Pre-Pilot Frame

Dependent variable	Pilot weeks 1 to 5 (1)	Pilot weeks 6 to 10 (2)	Experiment Days 1 to 20 (3)	Full rollout weeks 1 to 3 (4)	Full rollout later weeks (5)
Panel A. *ANCOVA: rank, pre-pilot outcome and pre-pilot covariates*					
Premium per day (yuan)	536	1,076*	3,632**	2,924	3,859**
	(512)	–	(1,810)	(2,000)	(1,790)
RI p	[0.287]	[0.054]	[0.046]	[0.134]	[0.012]
Non-pilot mean	5,180	5,918	15,440	20,190	16,850
Log premium per day	0.101	0.302**	0.194	0.168	0.361
	(0.115)	(0.151)	(0.212)	(0.224)	(0.236)
RI p	[0.391]	[0.072]	[0.431]	[0.493]	[0.114]
Conversions per day	0.004	0.025	0.059*	0.049	0.063*
	(0.014)	(0.021)	(0.032)	(0.045)	(0.037)
RI p	[0.810]	[0.224]	[0.064]	[0.263]	[0.118]
Non-pilot mean	0.42	0.45	0.59	0.81	0.69
Customers opened per day	0.43	0.57	1.07*	0.97	1.06
	(0.27)	(0.46)	(0.62)	(0.79)	(0.86)
RI p	[0.174]	[0.224]	[0.096]	[0.257]	[0.234]
Non-pilot mean	21.0	19.3	18.5	21.5	22.9
Active-day share	0.008	0.011	0.015	0.009	0.028
	(0.006)	(0.015)	(0.019)	(0.023)	(0.024)
RI p	[0.419]	[0.521]	[0.463]	[0.689]	[0.265]
Panel B. *Rank fixed effects only: premium per day (yuan) by frame*					
Pre-pilot frame (862 brokers)	670	1,227*	3,907**	3,438*	4,349**
	(533)	(652)	(1,850)	(2,090)	(1,810)
Active in pilot week 1 (858 brokers)	716	1,303**	4,073**	3,551*	4,504**
	(538)	(647)	(1,860)	(2,100)	(1,810)
Eligible roster (1,074 brokers)	1,224**	1,887***	6,015***	6,270***	6,565***
	(495)	(592)	(1,680)	(1,910)	(1,660)

Notes: Pilot minus non-pilot brokers, per calendar day (inactive days count as zero). Pre-pilot frame: brokers of ranks L2 to L5 on the roster at the start of the pilot who worked on at least three of the nine days before it; 862 brokers, 187 pilot brokers. Panel A adds to rank fixed effects the outcome's pre-pilot value and the pre-pilot covariates of Table 7. Panel B compares frames with rank fixed effects only: the pre-pilot frame; brokers active on at least three days of the first pilot week (858 brokers, 187 pilot brokers), a frame selected on a pilot outcome; and the full eligible roster (1,074 brokers, 204 pilot brokers), whose gap includes the effect of the pilot on whether a broker works at all. Robust standard errors in parentheses; the standard error of the premium estimate in column (2) of Panel A was not recovered from the exported output. Brackets: randomization-inference p-values from 500 permutations of pilot status within rank. Premium is winsorized at the 99th percentile of positive broker-days. Full rollout weeks 1 to 3 are the first three weeks after the experiment; later weeks run to the end of the observation window. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

6.4 Effort and Revenue per Unit of Effort

Does the advantage come from working more or from earning more for the same work? Appendix Table A43 answers in proportional terms, using the same frame and adjustment with Poisson estimates. Pooled from pilot week 6 to the end of the window, pilot brokers' premium is 15.0 percent higher (randomization-inference $p = 0.023$). Their recorded effort is only slightly higher: 1.5 percent more work days ($p = 0.56$), 4.0 percent more customers opened ($p = 0.18$), 4.5 percent more chat rounds ($p = 0.27$), 4.3 percent more telephone time ($p = 0.23$) and 6.3 percent more chat time ($p = 0.090$). None of these gaps changes at week 6, when the premium gap appears.

Premium per unit of recorded effort therefore remains higher. Per work day the gap is 13.1 percent ($p = 0.047$); per chat round 10.8 percent ($p = 0.050$); per customer opened 10.7 percent ($p = 0.056$); per minute of chat 7.9 percent ($p = 0.13$); and with all five effort measures as controls 8.2 percent ($p = 0.090$). Conversions per customer opened do not differ. Splitting the log premium gap of 0.14 into customers opened, conversions per customer opened and premium per conversion gives 0.040, 0.028 and 0.076: about half of the gap, and the only part significant at the 10 percent level, is larger purchases. This matches the customer experiment, where access raises premium mainly through life policies of 10,000 yuan or more (Section 3.3).

Two qualifications apply. Effort is recorded after assignment and may itself respond to access, so the per-unit gaps describe revenue per unit of recorded work rather than a causal split; if access also raised effort, they understate what access added per unit of work. And the per-unit gaps are largest in pilot weeks 6 to 10 and the experiment (16 to 20 percent per work day) and smaller in full rollout (about 10 percent, not statistically significant), while in the nine days before the pilot, premium per unit of effort is already 20 to 25 percent higher for pilot brokers, with standard errors of about 17 percentage points. Nine days cannot rule out a pre-existing difference in revenue per unit of effort.

6.5 Learning or Selection

The persistent residual is consistent with learning: brokers who used the tool longer may have learned to advise better, built routines, or accumulated returning customers who later bought, and the gap's delayed appearance fits such accumulation. The delay is weaker evidence than it seems, however, because the gap appears when demand for savings-type life insurance was rising, as guaranteed rates were expected to fall; if the tool's value lies in such sales, the gap could appear then even without learning. But the same pattern fits selection that the adjustment misses: pilot brokers worked more before the pilot, and a constant difference in unobserved effort or in revenue per unit of effort, growing with seasonal demand, produces a persistent proportional gap. Recorded effort alone cannot carry this account, because after adjustment it differs by

only a few percent (Section 6.4).

The available moments cannot separate these accounts. A simulated broker experience stock that accumulates with AI-assisted conversations and depreciates, with a free selection term for pilot brokers, is not recovered from the pilot gaps, the weekly customer-level effects and the random dose of Treat customers: the 95 percent profile sets for the learning parameters cover the whole grid, and learning is roughly recovered only if no selection is imposed, a restriction the balance test rejects (Appendix E). The direct dose test is also uninformative: no effect of the excess number of Treat customers a broker previously served is significant at the 5 percent level, and the upper bounds allow large effects (Appendix Table A41 and Figure A19). Nor do Control customers of non-pilot brokers catch up with those of pilot brokers during the experiment, so there is no sign of growing spillover into conversations without AI. Recorded effort does not settle it either: after adjustment, pilot brokers work only a few percent more, and most of the premium gap remains per unit of recorded effort (Section 6.4), but the per-unit gap is imprecise in full rollout and a pre-pilot difference in revenue per unit of effort cannot be ruled out.

Whether initial gains persist is a question about learning and technology adoption. AI can disseminate expertise (Brynjolfsson et al., 2025), but use, retained skill and labor-market outcomes are different objects (Humlum and Vestergaard, 2025b,a), complementary organizational investment matters (Brynjolfsson et al., 2021; Babina et al., 2026), and current production differs from human-capital development (Emanuel et al., 2026). The pilot shows that earlier access is followed by a lasting revenue advantage once every broker has the tool, earned mostly as more premium per unit of recorded work and through larger purchases rather than through more work. Because the measured selection runs toward more effort among pilot brokers, the gap more likely overstates than understates the dynamic return to early access, and I do not read it as an estimate of retained skill.

6.6 What Full Deployment Implies

The customer experiment compares customers of the same broker while the broker's other customers lacked the tool, and the pilot compares brokers while other brokers lacked it. Neither contrast is the effect of giving the tool to everyone. Four results speak to full deployment.

First, the customer-level effect does not come from displacing Control customers. The random daily number of Treat customers a broker serves does not change Control customers' 15-day premium (6 yuan per additional Treat customer, standard error 23) or the broker's daily premium and conversions (Appendix Table A19). Adding the same-day spillover gives an implied full-coverage effect of 159 yuan (standard error

107), against a customer-level effect of 84 yuan in the same sample, so the customer contrast neither clearly understates nor clearly overstates the effect of deployment. Broker capacity does bind: on days with more first chats, Treat customers' premium is lower (Appendix Table A32), so the gain per conversation is likely smaller where brokers are busier.

Second, initial access does more than move purchases forward. Once every customer has the tool, customers first assigned to Treat still buy more: 0.22 percentage points in full rollout alone and 0.48 points cumulatively by full-rollout week 7 (Table A16). The premium gap, in contrast, is earned mostly in the experiment window; in full rollout alone it is 35 yuan (standard error 74) for all customers and 129 yuan (139) for returning non-buyers.

Third, common access does not erase differences across brokers. The pilot gap does not converge after every broker has the tool (Section 6.3), so brokers who received the tool earlier keep their advantage, whether through learning or selection (Section 6.5).

Fourth, the ability gradient measured in the experiment is an upper bound for deployment. Part of it reflects how brokers divide effort between groups when they know which customers have the tool (Section 4.4); under full deployment there is no Control group to divert effort from. In the pilot, where this margin was absent, the gradient in large life policies remains but the premium gradient is close to zero.

The design does not identify the aggregate effect on the platform. Full rollout has no Control group and coincides with seasonal changes in demand for life insurance, so I do not compare periods. Prices, commissions, insurers' product offers and competition from other channels lie outside the data; the shift from no purchase toward life insurance (Section 3.4) is an expansion of purchases through this platform, not necessarily of the insurance market.

7 Interpreting the Gains: Product and Relationship

The experiment and the pilot measure what access to the tool is worth; they do not say why the gains differ by product, by the customer's history with the broker and by broker ability. Three patterns from Sections 4 and 5 constrain any explanation. The gains are in life insurance and work mainly through premium per buyer. Higher-ability brokers, measured by their sales performance before the experiment, gain more among customers who had already chatted with them, mainly through large life policies, while for new customers the gain varies little with ability and in health insurance the tool narrows gaps across brokers. Stronger brokers adopt a smaller share of the suggestions they receive. I first show where the ability gap comes from and then propose an interpretation in which the tool supplies content and the customer weighs it by how far she trusts the

broker.

7.1 What Drives the Ability Gap

Table 9 decomposes the effect of access on premium among returning non-buyers by quintile of the main ability index. The top quintile's 15-day premium effect is 513 yuan per customer and the bottom quintile's 24 yuan; the gap of 490 yuan (bootstrap standard deviation 177) is about 29 percent from a higher purchase probability (140 yuan) and 71 percent from higher premium per buyer (349 yuan). Among new customers, 79 percent of the 92-yuan effect is premium per buyer.

On the purchase records with product categories, the top-minus-bottom gap of 832 yuan is mostly life insurance (737 yuan), and its clearest component is basket composition, 373 yuan (95 percent bootstrap interval 76 to 856). In the top quintile, access adds life buyers (1.17 percentage points, standard error 0.45) and leaves health buyers unchanged; in the bottom quintile it adds health buyers (0.47 points, 0.21). Customers rarely buy both lines, and without the tool the life share of a broker's policies varies little with ability (Section 4.5), so I read this component as the tool helping different brokers close different needs, not as brokers moving customers from one product to the other.

7.2 Content and Credibility

A customer arrives with a need, for protection or for savings, that neither the broker nor the tool changes. The two needs differ in what advice must achieve. Health insurance in the data is mostly short-term medical cover priced at a few hundred yuan; its terms can be checked, and customers who buy decide quickly (median one day after the first chat). Life insurance in the data is mostly savings-type cover: the median policy is about 20 times a health policy, one life policy in ten is 100,000 yuan or more, the money is committed for years, and customers take longer to decide (median five days) (Appendix Table A1). Whether such a policy suits the customer is revealed only over years, so it is a credence good in the sense of Darby and Karni (1973): the customer must rely on the broker's word.

Let a broker of ability z supply content $k(z)$, the product facts, comparisons and illustrations in her replies, and credibility $r(z)$, how far the customer believes what she says. The tool adds content a to the conversations in which it is used, at a dose $\omega(z)$ that falls with ability, since stronger brokers adopt a smaller share of the suggestions they receive. For protection, the customer buys if the content resolves her questions,

$$P_P(z) = F(k(z) + \omega(z)a), \tag{5}$$

Table 9: Where the Ability Gap Comes From: Purchase Probability, Premium per Buyer and Products

	Returning non-buyers				New
	Bottom quintile (1)	Top quintile (2)	Top minus bottom (3)	Top minus bottom, linear (4)	customers (5)
Panel A. *Platform 15-day premium effect* (*yuan per customer*)					
Total	23.6	513.3***	489.7***	511.2***	91.7**
	(87.0)	(149.9)	(176.9)	(175.5)	(44.2)
From purchase probability	−43.4	97.0*	140.4**	115.3**	18.9
	(38.7)	(50.2)	(59.8)	(57.7)	(14.5)
From premium per buyer	66.9	416.3***	349.4**	395.9**	72.7*
	(72.9)	(144.5)	(163.3)	(166.0)	(42.4)
Share from premium per buyer (%)	–	81	71	77	79
Panel B. *Purchase records, days 0 to 14* (*yuan per customer*)					
Total	41.4	873.9*	832.4	862.3	43.8
	(343.5)	(496.7)	(588.3)	(568.9)	(138.5)
From purchase probability	67.7	270.8	203.1	129.5	76.3*
	(133.8)	(187.1)	(228.0)	(215.5)	(43.8)
From basket composition	−91.6	281.7	373.2*	401.8**	60.1*
	(63.0)	(184.7)	(195.5)	(190.9)	(33.4)
From premium within basket	65.3	321.4	256.1	331.0	−92.5
	(333.5)	(561.2)	(634.4)	(612.1)	(139.9)
Panel C. *Purchase records by product* (*yuan per customer*)					
Life insurance	49.0	786.4	737.4	777.1	33.7
	(343.9)	(489.9)	(583.8)	(562.3)	(138.9)
Health insurance	−0.8	50.0	50.8	42.0	7.4
	(10.3)	(43.9)	(45.4)	(46.5)	(4.8)
Other	−6.8	37.5	44.3	43.2	2.7
	(29.4)	(49.6)	(57.9)	(57.1)	(6.6)
Customers	10,096	9,964			68,397

Notes: Customers of non-pilot brokers with an ability index (120,586 customers, 746 brokers); returning non-buyers are split into quintiles of the main ability index. Every entry is an effect of access in yuan per customer, not a group mean. *Total* rows are regression coefficients: one OLS regression of the outcome on Treat interacted with six cells (the five quintiles of returning non-buyers and new customers), with broker, first-chat-date and customer-rating fixed effects; the coefficient on Treat in a cell is that cell's effect. To split a total into parts, each cell gets a Control level equal to the cell mean minus the Treat share times the coefficient, and a Treat level equal to the Control level plus the coefficient. The two levels differ by exactly the coefficient, so the parts add up to the total. Panel A splits the platform 15-day premium effect into a purchase-probability part and a premium-per-buyer part (Shapley decomposition of premium = purchase probability × premium per buyer). Panel B uses purchase records with product category on days 0 to 14 (capped at the 99th percentile) and splits the effect into purchase probability, the mix of baskets (life only, health only, other combinations) and premium within basket; Panel C splits the same total by product. Column (3) is column (2) minus column (1); column (4) replaces the cell levels by a linear fit across the five quintiles before taking the difference. The quintile effects differ from Appendix Table A22, which estimates each quintile separately on the larger index-only sample. Broker-cluster bootstrap standard deviations (300 draws) in parentheses; stars from the ratio of the estimate to the bootstrap standard deviation. The decomposition is additive, so it is in yuan by construction; proportional effects are in Table 5 and Appendix Table A22. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

with F increasing and concave. For savings, what counts is content she believes: as in Gennaioli et al. (2015), trust lowers the perceived risk of the product, so her decision to buy and the amount she commits rise with

$$\tau(z) = \tau_0 + r(z)[k(z) + \omega(z)a]. \tag{6}$$

Four implications follow. First, in protection the gain $F(k + \omega a) - F(k)$ falls with k: the tool narrows gaps across brokers. Second, in savings the gain $r(z)\omega(z)a$ rises with ability whenever credibility rises faster than use falls, and it is largest where amounts are large. Third, credibility depends on the relationship. A new customer does not yet know the broker, so r is low and similar across brokers and the tool helps them about equally; a customer who has already chatted with the broker has formed a view, so r rises with ability and the tool's content is worth more to strong brokers. Fourth, because the savings gain runs through the broker's credibility, the tool does not save broker time, and its gain should shrink when the broker is busy. In the knowledge hierarchies of Garicano (2000) and Ferreira and Li (2026), verification is done by the worker who checks the AI; in advice the verification that binds is the customer's, who checks the broker.

7.3 Evidence and Limits

The first implication matches the health results: the health-insurance gradient is negative (-0.20 percentage points per standard deviation, standard error 0.09), and in the bottom ability quintile of returning non-buyers access adds health buyers. The second matches the profile of effects across policy sizes (Appendix F): among returning non-buyers the Treat effect relative to the Control level rises from 17 percent for any life policy to 46 percent for policies of 100,000 yuan or more, and the difference between the top and bottom ability quintiles rises from 62 percent of the Control level to 145 percent. The third matches the history split of Section 4.4: the ability gradient sits among customers who had already chatted with the same broker, and for new customers the top-minus-bottom difference is small or negative at every policy size. The fourth matches the absence of time savings (Section E.4).

Three findings limit this reading. Larger health policies, which include long-term critical-illness cover, show the same narrowing as small ones, so the line between the two kinds of advice follows the product category more than the length of cover. The number of earlier chat days with the same broker does not change the gradient, which fits a relationship threshold better than richer history for the tool to read, but the tool reads that history, so the two cannot be fully separated. And editing of suggestions before sending is rare and does not rise with ability, so the adoption records cannot show

whether strong brokers rewrite suggestions in their own voice or set them aside. I do not estimate the parameters of equations (5) and (6): the moments available determine the dose and the content of the tool only jointly, so the model interprets the patterns rather than measuring credibility. Part of the experimental premium gradient also reflects how brokers divide effort between customers with and without the tool (Section 4.4); the gradient in large life policies, which survives where brokers could not divide effort, is the part the model explains.

An accounting split of the premium effect into suggestion receipt, adoption rate and implied return per adoption agrees with Section 5: the lower adoption rate of stronger brokers works against their gain, and the implied return per adoption carries it (Appendix Table A44). In the model, the return per adoption is the credibility $r(z)$ that multiplies the tool's content.

7.4 Counterfactuals

Because the platform matches customers to brokers as good as randomly, the experimental cells by customer class and broker ability give values for counterfactual matches. Sending returning non-buyers to the strongest brokers first, under each quintile's observed capacity, gains 73 yuan per customer more with the tool than without it (95 percent band 22 to 125); since part of the cell gradient reflects the division of effort between groups, this is an upper bound. For the firm, premium per AI suggestion is about 40 yuan at mean ability. Life insurance is most of that premium, and single-premium savings policies pay a first-year commission of only about 2 to 3 percent; with assumed rates of 3 to 35 percent, brokerage revenue per suggestion is about 1 to 20 yuan, against inference costs of 0.002 to 0.3 yuan, so revenue exceeds inference cost by a factor of about 3 under the least favorable combination. The binding uncertainty is the premium effect and the commission rate, not the cost of inference. Appendix F reports these calculations, an implied full-coverage effect that adds same-day spillovers, and the sensitivity of each.

8 Conclusion

This paper uses a randomized field experiment at Ant Insurance, China's largest insurance brokerage, to study how generative AI changes financial advice. When brokers receive generative-AI reply suggestions, customers buy more life insurance and pay more premium. Higher-ability brokers, measured by their recent sales performance, gain more among customers who already know them, mainly through large life policies; for new customers the gains vary little with ability, and in health insurance the tool narrows gaps. Brokers who adopt more suggestions do not gain more.

The general lesson concerns when AI levels and when it amplifies. The same tool, used by the same brokers on the same days, narrows gaps where customers can check the advice themselves and widens them where customers must trust the broker, as in savings products whose suitability is revealed only over years and with customers who already know the broker. Whether AI substitutes for or complements expertise depends on the product and the relationship, not only on the technology; evidence of leveling in customer support and writing and of amplification in advice are two parts of one pattern.

The consequences reach beyond the first conversation. Customers with access to AI advice return to chat and buy again more often, including health insurance later on, and pilot brokers' revenue advantage persists after every broker has the tool, mostly as more premium per unit of recorded work rather than more work; the customer-level effect does not come from displacing customers without the tool, and the aggregate effect on the platform is not identified (Section 6.6). For the firm, the premium gained per AI suggestion implies brokerage revenue of roughly 1 to 20 yuan per suggestion under assumed commission rates, against an inference cost of at most about 0.3 yuan.

Four limits qualify these conclusions. First, brokers knew which customers had the tool, and part of the experimental premium gradient reflects how they divided effort between the two groups; the gradient in large life policies remains in the pilot, where they could not, but the premium gradient does not, so the experimental premium gradient is an upper bound for full deployment. Second, pilot brokers worked more before the pilot, so the persistent pilot gap cannot separate learning from residual selection on unobserved effort or skill, even though recorded effort explains little of it. Third, the evidence comes from one platform, one tool and one market; the balance between leveling and amplification will likely differ with the product mix and the accuracy of the tool. Fourth, additional premium is not the same as better coverage: the extra life purchases are mostly savings-type policies that lock money in for years, so their welfare value depends on how well they fit household needs and liquidity.

For organizations that deploy AI, the results point to measuring realized customer outcomes rather than adoption, and to recognizing that experiments in which workers

know who has the tool can overstate how much it widens performance gaps. Where advice requires the customer's trust, the return to the broker's credibility rises.

References

Acemoglu, Daron and Pascual Restrepo, "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, 2018, *108* (6), 1488–1542.

_ **and** _ , "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives*, 2019, *33* (2), 3–30.

Agarwal, Nikhil, Alex Moehring, and Alexander Wolitzky, "Designing Human-AI Collaboration: A Sufficient-Statistic Approach," *NBER Working Paper 33949*, 2025.

_ **,** _ **, Pranav Rajpurkar, and Tobias Salz**, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," *NBER Working Paper 31422*, 2023.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb, "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction," *Journal of Economic Perspectives*, 2019, *33* (2), 31–50.

Anagol, Santosh, Shawn Cole, and Shayak Sarkar, "Understanding the Advice of Commissions-Motivated Agents: Evidence from the Indian Life Insurance Market," *Review of Economics and Statistics*, 2017, *99* (1), 1–15.

Angelova, Victoria, Will S. Dobbie, and Crystal S. Yang, "Algorithmic Recommendations and Human Discretion," *Review of Economic Studies*, 2026, *93* (4), 2250–2283.

Autor, David H., "Applying AI to Rebuild Middle Class Jobs," *NBER Working Paper 32140*, 2024.

Babina, Tania, "Understanding Firms' AI Efforts and Their Economic Impact," *The Review of Corporate Finance Studies*, 2026, *15* (4), 967–1013.

_ **, Alex X. He, and Renhao Jiang**, "Canaries in the Gold Mine: Early Productivity Gains from Artificial Intelligence Creating Organization Capital," Working Paper 35684, National Bureau of Economic Research 2026.

Berg, Tobias, Valentin Burg, Ana Gombović, and Manju Puri, "On the Rise of FinTechs: Credit Scoring Using Digital Footprints," *The Review of Financial Studies*, 2020, *33* (7), 2845–2897.

Bergstresser, Daniel, John M. R. Chalmers, and Peter Tufano, "Assessing the Costs and Benefits of Brokers in the Mutual Fund Industry," *Review of Financial Studies*, 2009, *22* (10), 4129–4156.

Bhattacharya, Vivek, Gastón Illanes, and Manisha Padi, "Fiduciary Duty and the Market for Financial Advice," *Econometrica*, 2025, *93* (4), 1449–1480.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, 2021, *13* (1), 333–372.

_ , Danielle Li, and Lindsey R. Raymond, "Generative AI at Work," *Quarterly Journal of Economics*, 2025, *140* (2), 889–942. First published 2023 as NBER Working Paper 31161.

Cai, Jing, Alain de Janvry, and Elisabeth Sadoulet, "Social Networks and the Decision to Insure," *American Economic Journal: Applied Economics*, 2015, *7* (2), 81–108.

_ , _ , and _ , "Subsidy Policies and Insurance Demand," *American Economic Review*, 2020, *110* (8), 2422–2453.

Charoenwong, Ben, Alan Kwan, and Tarik Umar, "Does Regulatory Jurisdiction Affect the Quality of Investment-Adviser Regulation?," *American Economic Review*, 2019, *109* (10), 3681–3712.

Choukhmane, Taha, Tim de Silva, Weidong Lin, and Matthew Akuzawa, "AI Financial Advice: Supply, Demand, and Life Cycle Implications," *MIT Sloan Working Paper 7377-26*, 2026. Earlier version circulated as "How Good is Generative AI Personal Financial Advice?".

Christoffersen, Susan E. K., Richard Evans, and David K. Musto, "What Do Consumers' Fund Flows Maximize? Evidence from Their Brokers' Incentives," *Journal of Finance*, 2013, *68* (1), 201–235.

Cong, Lin William, Xiang Hui, and Yushan Zhou, "AI Copilots in Real Estate: Evidence from China," *SSRN Working Paper 5498238*, 2025.

Cui, Kevin Zheyuan, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz, "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers," *Management Science*, 2026. Published online February 27, 2026.

D'Acunto, Francesco, Nagpurnanand Prabhala, and Alberto G. Rossi, "The Promises and Pitfalls of Robo-Advising," *Review of Financial Studies*, 2019, *32* (5), 1983–2020.

Darby, Michael R. and Edi Karni, "Free Competition and the Optimal Amount of Fraud," *Journal of Law and Economics*, 1973, *16* (1), 67–88.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality," *Organization Science*, 2026, *37* (2), 403–423.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114–126.

_ , _ , and _ , "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Management Science*, 2018, *64* (3), 1155–1170.

Egan, Mark, "Brokers versus Retail Investors: Conflicting Interests and Dominated Products," *Journal of Finance*, 2019, *74* (3), 1217–1260.

_ , Gregor Matvos, and Amit Seru, "The Market for Financial Adviser Misconduct," *Journal of Political Economy*, 2019, *127* (1), 233–295.

_ , Shan Ge, and Johnny Tang, "Conflicting Interests and the Effect of Fiduciary Duty: Evidence from Variable Annuities," *Review of Financial Studies*, 2022, *35* (12), 5334–5386.

Eisfeldt, Andrea L. and Gregor Schubert, "Generative AI and Finance," *Annual Review of Financial Economics*, 2025, *17*, 363–393.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs are GPTs: Labor Market Impact Potential of LLMs," *Science*, 2024, *384* (6702), 1306–1308.

Emanuel, Natalia, Emma Harrington, and Amanda Pallais, "The Power of Proximity to Coworkers," *Quarterly Journal of Economics*, 2026, *141* (3), 1825–1870.

Ferreira, Daniel and Jin Li, "Artificial Intelligence in the Boardroom," Technical Report, ECGI Finance Working Paper No. 1087/2025 2026.

Fieberg, Christian, Lars Hornuf, Maximilian Meiler, and David Streich, "Using Large Language Models for Financial Advice," *CESifo Working Paper 11666*, 2025.

Financial Stability Board, "Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Final Report," Technical Report, Financial Stability Board October 2026.

Foerster, Stephen, Juhani T. Linnainmaa, Brian T. Melzer, and Alessandro Previtero, "Retail Financial Advice: Does One Size Fit All?," *Journal of Finance*, 2017, *72* (4), 1441–1482.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *The Journal of Finance*, 2022, *77* (1), 5–47.

Garicano, Luis, "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 2000, *108* (5), 874–904.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, "Money Doctors," *Journal of Finance*, 2015, *70* (1), 91–114.

Gillen, Ben, Erik Snowberg, and Leeat Yariv, "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study," *Journal of Political Economy*, 2019, *127* (4), 1826–1863.

Goldstein, Itay, Wei Jiang, and G. Andrew Karolyi, "To FinTech and Beyond," *The Review of Financial Studies*, 2019, *32* (5), 1647–1661.

Greig, Fiona, Tarun Ramadorai, Alberto G. Rossi, Stephen P. Utkus, and Ansgar Walther, "Human Financial Advice in the Age of Automation," *Working Paper*, 2025. SSRN 4301514; revised July 7, 2025.

He, Longjiang, Zhenhua Li, Jinfei Sheng, Ying Yu, and Xiaoyan Zhang, "Does Your AI Move Your Money? Adoption and Impact of a Generative AI Investment Assistant," Working Paper, Tsinghua University PBC School of Finance 2026.

Hoffmann, Manuel, Sam Boysel, Frank Nagle, Sida Peng, and Kevin Xu, "Generative AI and the Nature of Work," *CESifo Working Paper 11479*, 2024.

Humlum, Anders and Emilie Vestergaard, "Still Waters, Rapid Currents: AI and the Structure of Modern Work," *NBER Working Paper 33777*, 2025.

_ **and** _ , "The Unequal Adoption of ChatGPT Exacerbates Existing Inequalities among Workers," *Proceedings of the National Academy of Sciences*, 2025, *122* (1), e2414972121.

Inderst, Roman and Marco Ottaviani, "Financial Advice," *Journal of Economic Literature*, 2012, *50* (2), 494–512.

Jiang, Jiajun, Miao Jin, Qin Li, Yu-Jane Liu, and Juanjuan Meng, "Behavioral Personalization and Algorithmic Explainability: Evidence from a Field Experiment on Robo-Advising," *SSRN Working Paper 5701864*, 2025.

Kanazawa, Kyogo, Daiji Kawaguchi, Hitoshi Shigeoka, and Yasutora Watanabe, "AI, Skill, and Productivity: The Case of Taxi Drivers," *Management Science*, 2026, *72* (2), 1376–1388.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

Koijen, Ralph S. J., Stijn Van Nieuwerburgh, and Motohiro Yogo, "Health and Mortality Delta: Assessing the Welfare Cost of Household Insurance Choice," *Journal of Finance*, 2016, *71* (2), 957–1010.

Linnainmaa, Juhani T., Brian T. Melzer, and Alessandro Previtero, "The Misguided Beliefs of Financial Advisors," *Journal of Finance*, 2021, *76* (2), 587–621.

Liu, Xing, "Artificial Intelligence and Customer Acquisition in Retail Financial Services: Experimental Evidence from Insurance Distribution," *Working Paper, Tsinghua University PBC School of Finance*, 2026. April.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore, "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes*, 2019, *151*, 90–103.

Mullainathan, Sendhil, Markus Noeth, and Antoinette Schoar, "The Market for Financial Advice: An Audit Study," *NBER Working Paper 17929*, 2012.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381*, 187–192.

Otis, Nicholas G., Rowan Clarke, Solène Delecourt, David Holtz, and Rembrand Koning, "The Uneven Impact of Generative Artificial Intelligence on Entrepreneurial Performance: Evidence from a Field Experiment in Kenya," *Management Science*, 2026. Published online July 10, 2026.

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

Reher, Michael and Stanislav Sokolinski, "Robo Advisors and Access to Wealth Management," *Journal of Financial Economics*, 2024, *155*, 103829.

Rossi, Alberto G. and Stephen P. Utkus, "The Diversification and Welfare Effects of Robo-Advising," *Journal of Financial Economics*, 2024, *157*, 103869.

Salvi, Francesco, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West, "On the Conversational Persuasiveness of GPT-4," *Nature Human Behaviour*, 2025, *9* (8), 1645–1653.

Silva, J. M. C. Santos and Silvana Tenreyro, "The Log of Gravity," *Review of Economics and Statistics*, 2006, *88* (4), 641–658.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone, "When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis," *Nature Human Behaviour*, 2024, *8*, 2293–2303.

Yang, Cathy L., Kevin Bauer, Xitong Li, and Oliver Hinz, "My Advisor, Her AI, and Me: Evidence from a Field Experiment on Human-AI Collaboration and Investment Decisions," *Management Science*, 2026, *72* (1), 242–264.

Appendix

Appendix A describes the setting, the data and the samples (Section 2). Appendix B reports robustness, inference and further purchase results (Section 3). Appendix C documents the ability index and the ability gradient (Section 4), and Appendix D the adoption results (Section 5). Appendix E covers the pilot (Section 6) and Appendix F the model (Section 7).

A Setting, Data and Sample

This appendix adds detail on the setting, the data and the samples of Section 2.

A.1 Market, Brokers and the Tool

Market and products. Intermediated premium in China was 962 billion RMB in 2024 across 2,162 licensed firms (*Yearbook of China's Insurance*, 2025). In 2024 Ant Insurance generated 9.43 billion RMB in revenue, approximately 5.9 percent of the intermediary market and roughly equal to the second and third largest firms combined, and about 694 million RMB in profit, roughly 15 percent of industry profit; its estimated premium volume is 55 to 70 billion RMB. The platform offers 1,264 products spanning health, life, accident, annuity, property, auto, travel, and pet lines. Life insurance specifies coverage amounts, policy terms, and beneficiary designations; critical-illness insurance additionally requires health disclosure, covers a defined list of diseases, and is subject to exclusion clauses and claim triggers that vary substantially across products. In the purchase records, the health category combines critical-illness and medical insurance and cannot separate them; its median premium of a few hundred yuan suggests that short-term medical cover dominates the count. Property insurance (including non-auto accident cover) and pet insurance are a negligible share of brokers' sales. Customers raise questions about coverage terms, health disclosure, premiums, and beneficiaries before deciding.

Broker rank and tenure. The rank follows the platform's internal career system and is determined by cumulative performance, training completion, and tenure. Among the non-pilot brokers in the main sample, L1, L2, L3, L4, and L5 account for 13.6, 56.4, 19.1, 8.5, and 1.7 percent. The median tenure of about seven months at the start of the pilot reflects that the platform was launched roughly three years before the experiment and is growing rapidly.

Table A1: Life and Health Insurance Compared

	Life	Health
Panel A. *Policies and premium*		
Policies	29,486	16,286
Share of policies (%)	57.7	31.9
Share of premium (%)	90.9	3.9
Policies with zero premium (%)	18.2	14.7
Premium per policy, mean (yuan)	32,346	2,394
Premium per policy, mean winsorized at P99 (yuan)	29,597	1,986
Premium per policy, P25 (yuan)	965	144
Premium per policy, median (yuan)	10,000	486
Premium per policy, P75 (yuan)	30,000	2,808
Premium per policy, P90 (yuan)	100,000	5,769
Premium per policy, P99 (yuan)	400,000	17,461
Premium per policy, P99.9 (yuan)	1,000,000	100,000
Policies of 10,000 yuan or more (%)	51.5	3.5
Policies of 100,000 yuan or more (%)	10.10	0.17
their share of the line's premium (%)	61.9	13.7
Panel B. *Buyers*		
Buyers	23,748	15,098
Age, mean (median)	37.1 (36)	34.5 (34)
Policies per buyer, mean	1.24	1.08
Buyers of this line only, share of all buyers (%)	54.2	33.5
Buyers who also bought the other line (%)	4.6	7.2
Panel C. *Path to purchase*		
Policies with a first chat in the previous 70 days	16,938	6,059
Bought on the first-chat day (%)	31.8	43.8
Bought within 7 days (%)	56.1	67.8
Bought within 30 days (%)	78.2	86.5
Days from first chat to purchase, median (mean)	5 (15.8)	1 (10.7)
Dialogue rounds to purchase, median (mean)	11 (23.0)	12 (23.7)
purchases through the third week of full rollout	15 (29.0)	17 (29.2)
Chat days to purchase, median (mean)	2 (2.9)	2 (2.5)
Dialogue rounds per chat day, median	4.8	6.0

Notes: Customers who chatted with any broker, pilot and non-pilot, from the first day of the pilot to the seventh week of full rollout. A policy is a customer, day and product line, with the largest premium recorded that day. Health insurance covers critical-illness and medical insurance, which the product field does not separate. Property, pet and policies without a product line (9.9 percent of policies and 5.1 percent of premium) are not shown. Premium is the amount paid for retained policies. Premium statistics in Panel A after the first four rows are over policies with positive premium. Panel C covers purchases from the transition days on and finds the first chat with any broker in the previous 70 days; days count from the first-chat day (day 0), and dialogue rounds and chat days count from the first chat through the purchase day. Instant-message records thin out late in the rollout window, so the indented row restricts rounds to purchases made by the third week of full rollout.

The generative-AI tool. The generative-AI tool described in Section 2 is known inside Ant Insurance as *Tongtiandai*. Figure A1 shows how it produces suggestions; its language model is based on Qwen2.5-72B, and the knowledge base covers product terms and regulatory requirements. Figure A3 shows the broker workstation with the tool active. The platform records each suggestion in one of seven states: fully adopted and sent; partially adopted and sent directly; partially adopted and sent after edits; copied as inspiration without sending in this conversation; not adopted; blocked by the platform's quality inspection; and exception. Exposure records receipt in the broker's workspace; adoption records the broker's subsequent handling. Access is enabled throughout Treat conversations, but generation takes time: many single-interaction conversations and some two-interaction conversations end before the tool can produce a reply, so zero exposure need not mean that a broker chose not to use AI. Across the experiment, about 3 percent of the suggestions non-pilot brokers received were fully adopted and between 6 and 9 percent were partially adopted and sent directly, depending on the week.

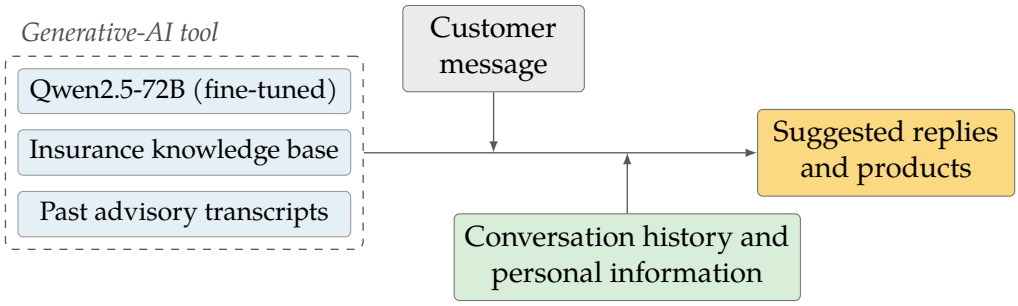


Figure A1: How the Generative-AI Tool Produces Suggestions for the Broker

Notes: For each incoming customer message, the generative-AI tool draws on a fine-tuned large language model, an insurance knowledge base, and a corpus of past advisory transcripts. It conditions on the customer's message, conversation history and personal information (age, gender, city, family composition, past insurance activity) and returns suggested replies and candidate products to the broker. Generation takes time: a short conversation can end before a suggestion is produced. The customer never sees the tool; all messages reaching the customer are sent by the broker.

Fields produced by the tool. For the 20 days of the experiment, the customer records hold, for each customer, broker and day, the number of AI suggestions the broker received for that customer and the number adopted, and a broker panel holds, for each broker and day, the number of AI suggestions and the number in each of the seven states: fully adopted, partially adopted and sent directly, partially adopted after editing, copied, not adopted, blocked by quality inspection, and exception. Every measure of exposure and adoption in the paper comes from these fields.

Interfaces. Figure A2 shows the screens through which a customer meets a broker; and Figure A3 the broker's workstation with the tool active. Table A3 defines every variable and Table A5 gives the steps from the raw records to the sample.



(a) Product page (b) Chat window (c) Broker profile (d) My insurance

Figure A2: Customer-Facing Interface: How a Customer Meets a Broker

Notes: Screenshots from the app. (a) A product page listing coverage terms and premiums, with an option to contact a licensed broker. (b) The chat window in which the advisory conversation takes place; in Treat conversations the tool is active on the broker's side, and AI suggestions reach the customer only if the broker sends them. (c) The broker's profile page. (d) The customer's page summarizing existing policies. No customer data are displayed.

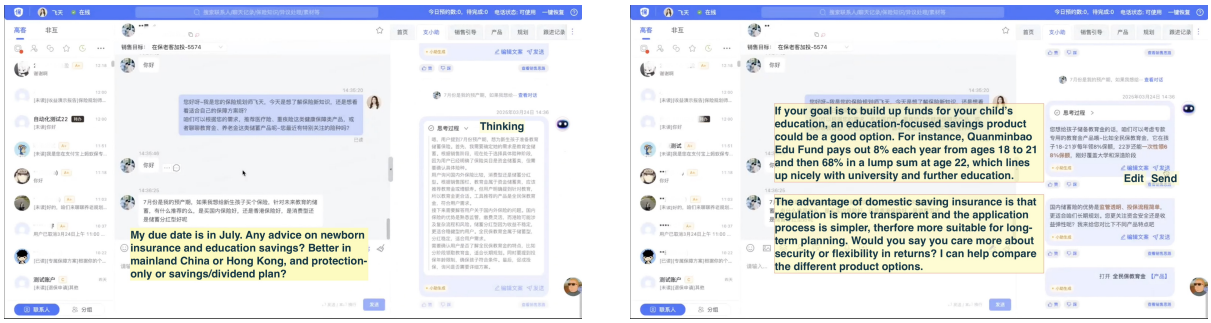


Figure A3: Broker Workstation with the Generative-AI Tool Active

Notes: Screenshots from the broker's workstation. Left: on receiving a customer message, the generative-AI tool processes it in real time. Right: the tool returns suggested replies in a side panel visible only to the broker, who decides whether to send, edit or ignore each one.

A.2 Data, Variables and Samples

Datasets. *Customer-broker-day records.* The primary dataset records, for each customer, broker and day on which the customer visits the broker, the daily count of dialogue rounds, same-day purchase and premium with the product category of each purchase (life, health, property including non-auto accident, pet), and the platform's 15-day purchase and premium fields. The platform computes the 15-day fields from its transaction records, which extend past the end of each window. For the 20 days of the experiment, the records also hold the group label, the platform's customer rating, the customer's age, and the number of AI suggestions the broker received and adopted for that customer. From the first day of the pilot to the seventh week of full rollout, 1,471,207 customers chatted with a broker at least once, and they paid 848.9 million yuan of premium; the records hold about 52,000 customer-broker records a day. I use the experiment records to define the experiment sample, the group label, AI exposure and adoption, and the records over the whole period for each customer's history before the first chat, for the first-chat date, for product categories, and for outcomes after the 15-day window.

Broker-day dataset. A daily panel of broker-level outcomes covers service time (IM and telephone duration), customers the broker opened a conversation with, conversion counts and premium, and life-insurance conversions and premium. I use it to measure broker ability before the experiment and for the pilot.

Broker static information and AI use. A broker roster gives each broker's rank, joining date (used to compute tenure) and a fixed indicator for pilot brokers. A broker panel for the experiment gives, for each broker and day, the number of AI suggestions the broker received, the number in each of the seven handling states, and the platform's ability score for the broker. Data from before the start of the pilot enter only the pilot analysis.

Outcomes built from daily purchase records. Outcomes by product category, by policy size and by day after the first chat come from the daily purchase records in days 0 to 14, with one record per customer and day for cumulative outcomes and one per customer, day and product category for product outcomes (the largest premium in each case). A product-coded purchase record is the policy unit in the product tables.

Samples. Table A2 summarizes the experiment samples (Panel A) and non-pilot brokers' output and effort in each phase in the first-pilot-week frame (Panel B). The main pilot frame and the first-pilot-week frame share 856 brokers; the 49 pilot brokers outside the roster frame are not in any pilot comparison. Within the main pilot frame the selection share is 19.5 percent in L2, 25.2 percent in L3, 20.7 percent in L4 and 44.0 percent in L5. The first-pilot-week frame conditions on activity after the pilot began.

Table A2: Analysis Sample

Panel A. Customers in the experiment

	Non-pilot brokers (main)		All brokers	
	Count	Share	Count	Share
Customers (engaged in ≥ 1 conversation)	124,347		164,732	
Assigned to Treat (AI enabled)	100,003	80.4%	132,996	80.7%
Assigned to Control	24,344	19.6%	31,736	19.3%
Brokers serving these customers	831		1,070	
15-day purchase rate, Control group		3.28%		3.37%
Treat customers with at least one AI suggestion		45.6%		46.3%
Control customers with at least one AI suggestion		0.7%		0.6%

Panel B. Brokers in the pilot frame: non-pilot brokers' means

	Pilot			Full rollout	
	Weeks 1 to 5	Weeks 6 to 10	Experiment	Weeks 1 to 3	Weeks 4 to 7
Per calendar day					
Premium (yuan)	5,211	5,957	15,540	20,310	16,990
Conversions	0.42	0.45	0.60	0.81	0.69
Customers opened	21.2	19.4	18.6	21.6	23.0
Share of days active	0.97	0.93	0.89	0.87	0.84
Effort					
Work days per week	6.79	6.48	6.25	6.10	5.91
Chat minutes per work day	59.3	62.0	67.2	77.1	71.1
Phone minutes per work day	36.4	36.3	35.2	41.1	47.5

Notes: Panel A: customers who chatted with a broker at least once during the experiment; group and broker from the first chat record. Shares for the group rows are relative to all customers in the column; the AI-suggestion rows are shares of each group whose broker received at least one AI suggestion for them. Panel B: the first-pilot-week frame, the 858 brokers of ranks L2 to L5 on the roster at the start of the pilot who worked on at least three days of its first week (187 pilot brokers; the main pilot frame is defined in Section 2); means over the non-pilot brokers in this frame, with inactive days counted as zero in the per-calendar-day rows. Premium per day is winsorized at the 99th percentile of positive values. Effort rows divide by the broker's work days (days with any recorded activity), which leaves 855 brokers in the frame.

Table A3: Variable Definitions

Variable	Definition	Source	Window
Panel A. *Assignment and sample*			
Treat	1 if the label on the customer's first chat record is Treat, 0 if Control	Customer records (experiment)	First chat record
Chat day	Customer-day with at least one dialogue round	Customer records	Any day
First-chat day	First chat day, from the first transition day on	Customer records	Transition days to Day 20
Broker	Broker on the first chat record	Customer records (experiment)	First-chat day
Pilot broker	1 if the broker had early access to AI advice in the pilot	Broker roster	Fixed
Panel B. *Customer characteristics (pre-treatment)*			
Prior funnel stage	Deepest stage reached with any broker: new customer, prior visitor, prior chatter, prior buyer	Customer records	Days −70 to −1
Returning non-buyer	1 if prior visitor or prior chatter	Customer records	Days −70 to −1
Same broker	1 if any record in the history window is with the first-chat broker	Customer records	Days −70 to −1
Days since last contact	Days from the last record with any broker to the first-chat day	Customer records	Days −70 to −1
Visit days; chat days	Distinct days with a record; with a chat	Customer records	Days −70 to −1
Prior premium	Premium of purchases before the first chat (one record per customer and day), winsorized	Customer records	Pilot period
Customer rating	Platform rating at first contact: B, A, S, S+, New customer	Customer records (experiment)	First contact
Age	Age in years	Customer records (experiment)	First contact
Panel C. *Broker characteristics and ability measures*			
Rank	Rank L1 to L5, or missing	Broker roster	Start of the pilot
Tenure	Months since hire	Broker roster	Start of the pilot
Comparison platform score	Platform ability score, sign flipped so that higher means stronger; standardized	Broker AI-use panel	Experiment
Ability index	First principal component of tenure, pre-experiment premium per day, conversions per day and conversions per opened customer; non-pilot brokers; standardized; pre-window premium cap	Daily records and roster	Before the customer experiment
Panel D. *Customer outcomes (post-treatment)*			
Purchase, 15 days	Platform 15-day conversion field on the first-chat-day record	Customer records	Days 0 to 14
Premium, 15 days	Platform 15-day premium field (total premium paid); winsorized at the 99th percentile of positive values	As above	Days 0 to 14
Purchase and premium, 30 days	15-day fields plus purchases on days 15 to 29 (one record per customer and day); lower bound	Customer records	Days 0 to 29
Same-day purchase	Platform purchase field on the first-chat-day record	As above	Day 0
Life, health policy; large life policy	Product-coded daily purchase record; large life: daily category premium of 100K yuan or more	Purchase records with product category	Days 0 to 14 (29)
Chat rounds; chat days; days with a record	log(1 + rounds); distinct chat days; distinct days with a record	Customer records	Days 0 to 14
Return to chat	1 if a chat on days 1 to 14 (or 1 to 3, 4 to 7, 8 to 14, 15 to 30)	Customer records	Days 1 to 30
Cumulative purchase and premium	Any purchase; sum of premium (one record per customer and day)	Customer records	To full-rollout week 3; to full-rollout week 7
Panel E. *AI use (post-treatment)*			
Received suggestions	1 if the broker received at least one AI suggestion while serving the customer	Customer records (experiment)	From the first-chat day
Adopted a suggestion	1 if the broker adopted at least one AI suggestion for the customer	Customer records (experiment)	From the first-chat day
Adoption states	Shares of suggestions fully adopted, partly adopted, not adopted, blocked by quality inspection	Broker AI-use panel	Experiment
Panel F. *Broker outcomes for the pilot*			
Premium, conversions, opened customers per day	Calendar-day averages, inactive days as zero; premium winsorized	Broker daily records	Pilot to end of data
Active share; phone and chat time	Share of days active; log(1 + seconds per day)	Broker daily records	As above

Notes: Day 0 is the first-chat day. Cumulative outcomes use one record per customer and day, and product outcomes one record per customer, day and category (the highest premium). Money is in yuan. The data record only policies that were not refunded.

Table A4: Summary Statistics of Outcome and Analysis Variables

Variable	*N*	Mean	SD	P1	P5	P10	P25	P50	P75	P90	P95	P99
Panel A. *Outcomes*												
Purchase on the first-chat day (%)	124,347	2.04										
Purchase, 15 days (%)	124,347	3.30										
Purchase, 30 days (%)	124,347	3.96										
Premium, 15 days, buyers (yuan)	4,099	11,090	29,780	0	0	0	100	678	6,000	30,000	60,000	200,000
Premium, 30 days, buyers (yuan)	4,929	15,580	41,420	0	0	0	114	1,000	10,000	50,000	100,000	250,000
Life policy, 15 days (%)	124,347	2.53										
Life premium, 15 days, life buyers (yuan)	3,143	40,800	90,730	0	0	0	500	4,485	40,000	100,000	200,000	502,500
Policies per buyer, 15 days	4,999	1.19	0.52	1	1	1	1	1	1	2	2	3
Life share of policies, buyers (%)	4,999	61.5	47.9	0	0	0	0	100	100	100	100	100
Purchase to full-rollout week 7 (%)	124,347	5.02										
Panel B. *Engagement, days 0 to 14*												
Chat rounds	124,347	5.46	10.96	1	1	1	1	3	5	11	19	46
Instant-message rounds	124,347	3.77	10.53	0	0	0	0	1	3	9	16	43
Chat days	124,347	1.35	0.79	1	1	1	1	1	1	2	3	5
Days with a record	124,347	1.72	1.23	1	1	1	1	1	2	3	4	7
Return to chat, days 8 to 14 (%)	124,347	8.69										
Chat rounds before purchase, buyers	4,099	11.52	15.06	1	2	3	3	6	13	26	38	71
Days to first purchase, buyers (days)	4,605	2.02	3.31	0	0	0	0	0	3	7	10	13
Panel C. *Customer characteristics before the first chat*												
Returning customer (%)	124,347	43.05										
Prior contact with the same broker (%)	124,347	34.76										
Visit days, 70 days before (days)	124,347	1.78	3.54	0	0	0	0	0	2	6	9	17
Chat days, 70 days before (days)	124,347	0.44	1.22	0	0	0	0	0	0	1	3	6
Days since last contact, returning (days)	53,528	14.9	18.4	1	1	1	2	6	23	46	58	68
Panel D. *AI use*												
Received an AI suggestion, Treat (%)	100,003	45.35										
AI suggestions, Treat	100,003	3.16	13.55	0	0	0	0	0	2	8	15	39
Adopted a suggestion, Treat (%)	100,003	19.49										
Adoptions, Treat	100,003	0.40	1.61	0	0	0	0	0	0	1	2	6
Broker share fully adopted (%)	816	3.40	4.07	0.0	0.0	0.3	1.0	2.5	4.5	7.0	8.8	18.1
Panel E. *Brokers*												
Tenure at the start of the pilot (months)	826	8.2	8.2	−2.6	−2.2	−1.4	2.9	6.7	10.8	20.5	22.7	39.2
Experiment customers per broker	831	149.6	66.1	1	6	57	119	153	188	228	258	306
Pilot: premium per day (yuan)	668	5,647	5,770	4	207	554	1,456	3,676	7,945	12,770	16,670	30,400
Pilot: opened customers per day	668	20.4	6.6	1.5	6.8	12.4	16.9	20.8	24.2	28.3	30.7	35.0
Pilot: instant-message minutes per working day	668	60	25	15	26	31	42	57	74	90	105	144

Notes: Customers of the 831 non-pilot brokers who chatted at least once in the experiment (124,347). Day 0 is the first-chat day. For 0/1 variables the mean is in percent. Premium is winsorized at the 99th percentile of positive values. Policies: purchase records with a product category, one per customer, day and category. Panel D: Treat customers, days 0 to 14; full-adoption share per broker over the experiment. Panel E: brokers of the main sample; tenure is negative for brokers hired after the start of the pilot. Pilot rows: non-pilot brokers of ranks L2 to L5 active in pilot week 1, weeks 1 to 10; premium per calendar day, winsorized.

Table A5: Sample Construction

Step	Customers	Brokers	Treat share
(1) Customers with an interaction record in the experiment	563,669	1,096	not meaningful
(2) Sent at least one message (experiment sample)	164,732	1,070	0.807
All brokers	164,732	1,070	0.807
New customers	94,032		0.810
Prior visitors	37,717		0.810
Prior chatters	30,666		0.797
Prior buyers	2,317		0.795
With the ability index (%)	73.2		
(3) Non-pilot brokers (main sample)	124,347	831	0.804
New customers	70,819		0.807
Prior visitors	27,942		0.806
Prior chatters	23,637		0.793
Prior buyers	1,949		0.795
With the ability index (%)	97.0		

Notes: The Treat share is reported from step (2) on, since assignment takes place at the first message. The group and the broker come from the customer's first chat record; the first-chat day can fall on a transition day. Prior funnel stage uses the 70 days before the first chat. The ability index covers non-pilot brokers only.

Customer rating. The rating is recorded on each customer record. It is distinct from the history-based classes of Section 2.5. Among Control customers of non-pilot brokers, B and A together account for 86.6 percent, S and S+ for 10.2 percent, and 3.1 percent are in the New customer category. The rating differs between the customer's first record in the experiment and the first chat record for only 0.41 percent of customers (0.41 percent in the Treat group, 0.39 percent in the Control group), and every rating category is balanced on both records.

History window. Seventy days before the earliest first chat is the first day of the pilot, so the history window has the same length for every customer and uses no data from before the pilot. Figure 2(b) shows the mix of first chats by day: new customers account for about half of first chats on every experiment day.

Treat share on the transition days. Among customers of non-pilot brokers the Treat share is 0.79 and 0.72 on the two transition days. First-chat-date fixed effects absorb this variation together with common calendar-day conditions when customers begin chatting.

A.3 Balance

Figure A4 plots the standardized differences between groups in the characteristics fixed before the first chat.

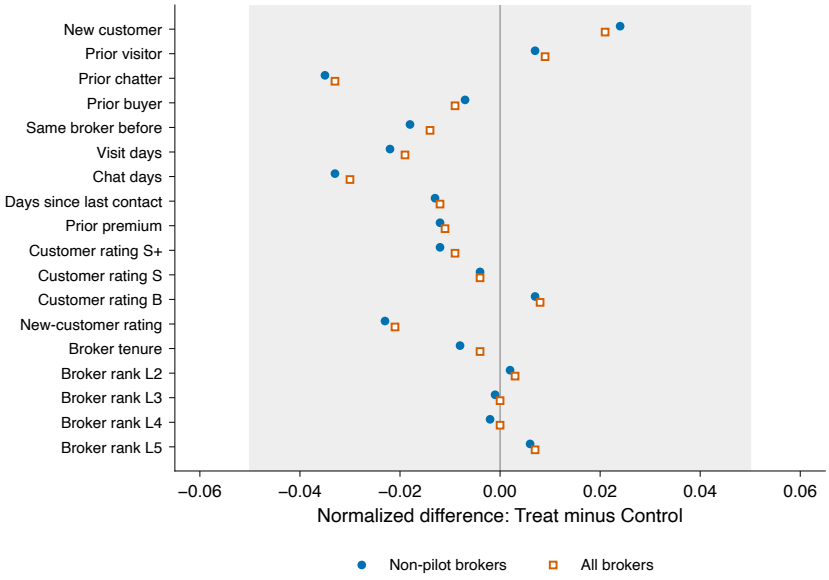


Figure A4: Balance at the First Chat

Notes: Normalized differences (Treat minus Control mean divided by the square root of the average group variance) of customer and broker characteristics fixed before the first chat, for non-pilot brokers (filled) and all brokers (hollow). The shaded band marks ± 0.05.

B Purchases: Additional Results

This appendix reports robustness checks, inference and further purchase results for Section 3.

B.1 Robustness and Inference

Table A6 reports the robustness checks discussed in Section 3.7, and Table A9 the earlier cell-statistic randomization-inference and Romano and Wolf p-values. Table A8 reports RI aligned with the main regression.

Table A6: Robustness of the Main Estimates

Dependent variable: Specification	Purchase, 15 days (pp) (1)	Premium, 15 days (yuan) (2)	Premium, PPML 15 days (%) (3)	Control mean (4)	Customers (5)
Panel A. *Non-pilot brokers*					
Baseline (Table 2)	0.12	88.4**	28.2**	315.4	124,347
	(0.13)	(40.2)	(15.5)		
Add prior funnel stage fixed effects	0.11	87.5**	29.2**	315.4	124,347
	(0.12)	(40.1)	(15.4)		
Premium winsorized at the 95th percentile		58.3**	24.0**	246.8	124,347
		(23.7)	(11.5)		
Premium winsorized at the 99.5th percentile		101.9**	29.8*	343.0	124,347
		(51.1)	(18.3)		
Drop customers whose first chat is on a transition day	0.13	85.4**	29.1**	294.6	119,386
	(0.12)	(40.3)	(16.7)		
Drop the transition days and Day 1	0.04	64.5	21.1	295.5	109,651
	(0.14)	(42.2)	(16.4)		
Panel B. *All brokers*					
Baseline (Table 2)	0.02	77.0**	22.5**	338.5	164,732
	(0.11)	(36.1)	(12.5)		
Add prior funnel stage fixed effects	0.03	77.4**	24.5**	338.5	164,732
	(0.11)	(36.1)	(12.6)		
Premium winsorized at the 95th percentile		42.2*	15.6*	267.6	164,732
		(21.9)	(9.3)		
Premium winsorized at the 99.5th percentile		82.2*	21.4	379.9	164,732
		(48.9)	(15.1)		
Drop customers whose first chat is on a transition day	0.02	74.2**	22.5*	322.3	158,707
	(0.11)	(36.2)	(13.2)		

Notes: OLS with broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. The Control mean refers to the premium outcome of the row. PPML column: Poisson pseudo-maximum likelihood with the same fixed effects, $100 \times (\exp(\beta) - 1)$ with delta-method standard errors; brokers, days or ratings with no premium drop out, so its samples are slightly smaller. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Premium caps. Table A7 reports the average premium effect under alternative caps. Effects are significant at the 5 percent level for caps between the 97.5th and the 99.5th percentile of positive premium at 15 days. With the 99.9th percentile or no cap, the point estimates stay between 89 and 102 yuan, but the standard errors almost double because of a few very large policies, and the estimates are not statistically significant.

Table A7: Premium Effects under Alternative Caps

Dependent variable: premium	All customers		Returning non-buyers	
Cap (percentile of positive premium)	15 days	30 days	15 days	30 days
Panel A. *OLS, yuan*				
97.5th	69.8**	120.6**	89.0**	199.0***
	(30.0)	(47.1)	(44.9)	(73.2)
99th (main)	88.4**	152.2**	113.0**	214.3**
	(40.2)	(59.6)	(57.4)	(100.6)
99.5th	101.9**	148.2*	104.3	166.0
	(51.1)	(76.4)	(81.7)	(141.6)
99.9th	89.1	114.9	59.4	79.8
	(70.8)	(98.6)	(130.7)	(198.1)
Uncapped	101.6	127.3	86.2	106.6
	(74.1)	(100.9)	(138.7)	(203.5)
Panel B. *PPML, percent change*				
97.5th	25.3**	25.3**	33.3*	42.2***
	(13.0)	(11.7)	(20.4)	(19.4)
99th (main)	28.2**	29.0**	39.0*	40.1**
	(15.5)	(13.7)	(24.4)	(23.0)
99.5th	29.8*	25.7*	31.6	27.4
	(18.3)	(15.7)	(29.6)	(25.8)
99.9th	23.3	18.3	17.0	13.3
	(22.0)	(17.9)	(35.8)	(28.6)
Uncapped	26.9	20.6	25.0	17.9
	(23.5)	(18.6)	(40.7)	(30.8)
Cap values (yuan), 97.5th / 99th / 99.5th / 99.9th	15 days: 102,600 / 200,000 / 414,000 / 1,000,000			
	30 days: 160,000 / 300,000 / 500,000 / 1,000,000			

Notes: Coefficient on Treat for premium (yuan) capped at the stated percentile of positive values. Customers of non-pilot brokers: 124,347 in all, 51,579 returning non-buyers. Broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. Panel B: Poisson pseudo-maximum likelihood with the same fixed effects, $100 \times (\exp(\beta) - 1)$ with delta-method standard errors; stars test the Poisson coefficient. The proportional effect has the same sensitivity to the upper tail as the effect in yuan. $^{***}\ p < 0.01$, $^{**}\ p < 0.05$, $^{*}\ p < 0.10$.

Threshold effects and the mean. Average capped premium is the integral of the exceedance probabilities in Figure 4 up to the cap; bracketing the step function between grid points gives an implied 15-day effect between 55 and 106 yuan, around the 88-yuan regression estimate, and nearly all of it comes from thresholds of 10,000 yuan and above. Above 100,000 yuan the threshold effect is 0.024 percentage points (standard error 0.019). Capping at the 97.5th, 99th and 99.5th percentiles of positive premium gives 70, 88 and 102 yuan; the uncapped mean is 102 yuan (standard error 74).

Randomization inference. The randomization-inference p-value of the 15-day premium coefficient is 0.064 (0.081 for the studentized statistic); for the 15-day purchase probability it is 0.34. For returning non-buyers the premium p-values lie between 0.06 and 0.08.

Table A8: Randomization Inference

Customers	Dependent variable	OLS effect	RI p_b	RI p_t	MCSE(p_b)	PPML (%)	RI p_s	Wild p_s
All non-pilot customers	Purchase, 15d	0.116	0.3377	0.3201	0.0067			
	Premium, 15d	88.370	0.0644	0.0808	0.0035	28.2	0.0846	0.0276
	Purchase, 30d	0.284	0.0412	0.0390	0.0028			
	Premium, 30d	152.200	0.0252	0.0282	0.0022	29.0	0.0282	0.0092
	Life purchase, 15d	0.331	0.0018	0.0014	0.0006			
Returning non-buyers	Purchase, 15d	0.110	0.5189	0.5143	0.0071			
	Premium, 15d	113.000	0.0616	0.0776	0.0034	39.0	0.0820	0.0455
	Purchase, 30d	0.396	0.0816	0.0814	0.0039			
	Premium, 30d	214.300	0.0620	0.0810	0.0034	40.1	0.0792	0.0237

Notes: Treat is permuted 4,999 times within broker × first-chat-date cells, preserving each cell's treated count. The sample and fixed effects are unchanged. Statistics are the additive-FE OLS coefficient and broker-cluster CR0 t. Tests are two-sided against the sharp null of no effect for any customer. Purchase effects are in pp and premium in yuan. MCSE is simulation uncertainty, not the sampling standard error. Samples: 124,347 and 51,579. Cell-FE inference in the earlier sensitivity table uses a different statistic. Last three columns: Poisson pseudo-maximum likelihood with broker, first-chat-date and customer-rating fixed effects ($100 \times (\exp(\beta) - 1)$); RI p_s permutes Treat in the same cells 4,999 times with the efficient score of the restricted Poisson model as the statistic; wild p_s is the score bootstrap with the null imposed (Rademacher weights by broker, 9,999 draws).

Two purchase definitions. Product, size, 30-day and follow-up outcomes use the daily purchase records; the primary outcomes use the platform's 15-day field. Over days 0 to 14 the two definitions overlap but are not identical: among customers of non-pilot brokers, 2.23 percent of the Treat group and 2.13 percent of the Control group buy under both, 1.13 and 1.20 percent only under the platform field, 1.88 and 1.83 percent only in the purchase records, and 94.77 and 94.84 percent under neither. The four shares are nearly the same in the two groups, so the difference between the definitions does not vary with assignment. Among customers who buy under both, 78 percent have the same premium in both.

Table A9: Earlier Cell-FE Inference and Family Adjustments

Hypothesis	Within-cell estimate	RI p	Romano-Wolf p
Panel A. *Randomization inference and Romano-Wolf*			
All customers: purchase, 15 days (pp)	0.093	0.530	0.775
All customers: premium, 15 days (yuan)	65.1	0.130	0.490
New customers: purchase, 15 days (pp)	0.137	0.495	0.775
New customers: premium, 15 days (yuan)	71.1	0.190	0.560
Returning non-buyers: purchase, 15 days (pp)	0.078	0.775	0.775
Returning non-buyers: premium, 15 days (yuan)	97.8	0.190	0.560
All customers: purchase, 30 days (pp)	0.248	0.135	
All customers: premium, 30 days (yuan)	129.9	0.045	
Panel B. *Benjamini-Hochberg q-values by prior funnel stage*			
	Purchase	Premium	Life policy
New customers	0.573	0.280	0.280
Prior visitors	0.739	0.280	0.908
Prior chatters	0.432	0.739	0.140
Prior buyers	0.973	0.973	0.573
Returning non-buyers	0.866	0.280	0.280
Panel C. *Primary family (common ability sample), OLS and PPML*			
	Statistic	Wild p	Holm p
M1 Treat, premium, 15 days: OLS (yuan)	93.1	0.039	0.044
M1, PPML (%)	29.3	0.025	0.066
M2 Treat × ability, premium, 15 days: OLS (yuan per SD)	106.8	0.010	0.030
M2, PPML (% per SD)	23.7	0.046	0.066
M3 Life minus health purchase gradient (pp per SD)	0.33	0.022	0.044 / 0.066

Notes: Archived sensitivity results using the cell-FE statistic. Main additive-FE randomization inference is reported separately in Table A8; these statistics and simulation runs are not interchangeable. Customers of non-pilot brokers. Panel A: randomization inference permutes Treat within broker by first-chat-day cells (200 draws; within-cell estimator); Romano-Wolf step-down p-values for the six 15-day hypotheses, from the same permutations. Panel B: Benjamini-Hochberg q-values across all class by outcome cells of Table A10, 15 days. Panel C: the three primary tests on 120,586 customers of 746 brokers. OLS: wild cluster bootstrap with the null imposed; PPML: Poisson pseudo-maximum likelihood with broker, first-chat-date and customer-rating fixed effects and the score bootstrap with the null imposed (Rademacher weights by broker). M3 is a purchase outcome and is the same in both; its Holm p is 0.044 in the OLS family and 0.066 in the PPML family.

B.2 Customer History and Customer Characteristics

Customer history. Figure A5(a) shows effects by prior funnel stage, the deepest stage a customer had reached with any broker in the ten weeks before the first chat, as a percent of each class's Control mean; Table A10 reports the estimates. The 95 percent interval for new customers' premium effect is 2 to 182 yuan on a Control mean of 240 yuan; returning non-buyers' Control mean is 297 yuan; over 30 days the effects are 128 and 214 yuan. The purchase effects are small in both classes. Equality of the premium effects across the four classes is not rejected ($p = 0.57$), and no Benjamini and Hochberg q-value falls below 0.14 (Table A9). Prior buyers are too few to estimate precisely. For returning non-buyers, purchases of life policies with premiums of at least 50,000 yuan rise by 0.21 percentage points within 15 days and 0.28 points within 30 days; new customers buy more small and middle-sized life policies (Figure A6). Figure A5(c) splits returning non-buyers by the time since their last contact with the broker: the premium effect is positive in every band, the purchase effect stays close to zero, and the intervals are wide. One explanation for larger effects among customers with past conversations is that the tool reads the customer's conversation history, so prior contact supplies context for its suggestions as well as a relationship with the broker (Cai et al., 2015, 2020). There is no statistically significant heterogeneity across customer ratings or age bands (Tables A12 and A11); among customers whose age is recorded no age band shows a significant effect, including the youngest and the oldest, though power in these small groups is low.

Table A12 and Figure A7 report effects by customer rating and age. Figure A8 reports purchase effects by prior funnel stage.

B.3 Products, Baskets and Premium per Buyer

Baskets and diversion. Table A13 reports the basket effects and the bounds on where induced life buyers come from. Without restrictions any inflow into life can be relabeled, so the shares are uninformative. Restricting transitions out of life, or ruling out lost purchases, narrows the mass of induced life buyers to 0.35 to 0.73 percentage points but leaves the shares from health and from no purchase almost unrestricted. Only the directional assumptions narrow the health share (at most 0.10 at the estimates), and only the last assumption gives point estimates.

Diversion by predicted health propensity, timing and later purchases. Table A14 reports effects by quintile of the predicted probability of a health purchase (Panel A) and on purchases of each product on days 15 to 45 (Panel B). Table A15 reports cumulative effects by product and day since the first chat, and the share of buyers who had bought by each day, by group.

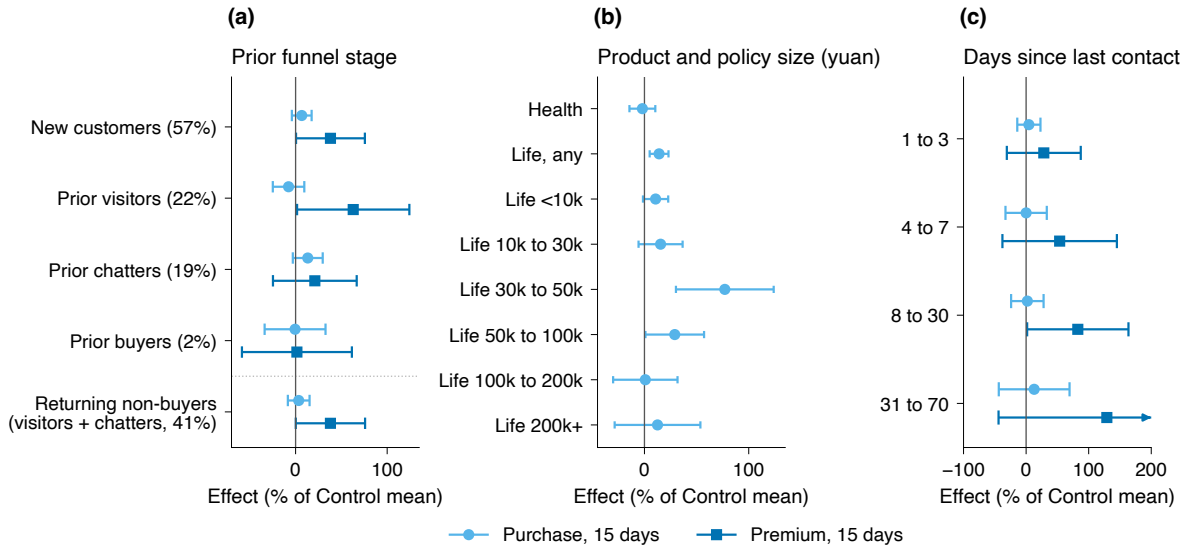


Figure A5: Effects by Customer History and Product

Notes: Customers of non-pilot brokers. Each marker is the coefficient on Treat as a percent of the Control mean, with a 95 percent confidence interval; arrows mark intervals that extend beyond the axis. Circles: purchase within 15 days; squares: 15-day premium. Panel (a): by prior funnel stage, with each class's share of customers in parentheses; returning non-buyers pool prior visitors and prior chatters. Panel (b): purchase by product and life-policy size (yuan), from purchase records. Panel (c): returning non-buyers, by days since their last contact with the broker. Broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker.

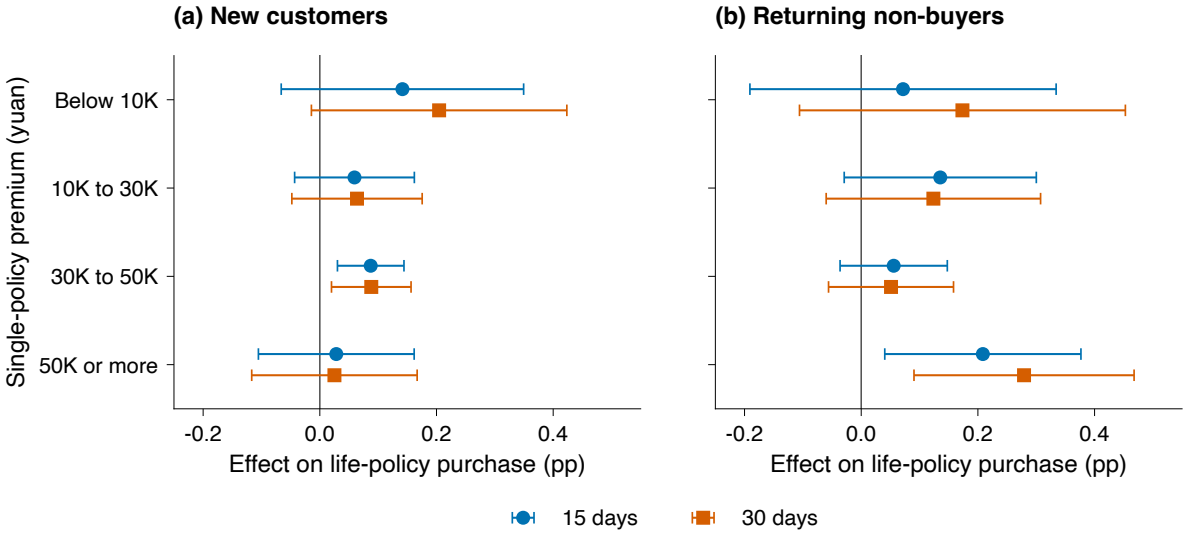


Figure A6: Life-Insurance Purchases by Customer History and Policy Size

Notes: Customers of non-pilot brokers. Treat effects on life-policy purchases by product-coded premium within 15 days (circles) and 30 days (squares). Panel (a): new customers; panel (b): returning non-buyers. K denotes 1,000 yuan of premium. Customers can buy in multiple bins. Common axes; 95 percent confidence intervals clustered by broker. Outcomes use product-coded purchase records (Appendix A), with broker, first-chat-day and customer-rating fixed effects.

Table A10: Effects by Prior Funnel Stage and Product

Dependent variable:	15 days		30 days		Product, 15 days			
	Purchase (pp) (1)	Premium (yuan) (2)	Purchase (pp) (3)	Premium (yuan) (4)	Life policy (pp) (5)	Health policy (pp) (6)	Life premium (yuan) (7)	Life policy ≥100k (pp) (8)
Panel A. *Non-pilot brokers*								
New customers	0.20	92**	0.30*	128**	0.26*	−0.04	36	−0.01
	(0.16)	(46)	(0.18)	(61)	(0.13)	(0.11)	(135)	(0.05)
Control mean	2.98	240	3.32	350	2.00	1.28	782	0.27
Prior visitors	−0.21	116**			0.07	−0.16	247	0.09
	(0.24)	(57)			(0.21)	(0.15)	(244)	(0.08)
Control mean	2.79	184			2.14	0.96	769	0.22
Prior chatters	0.50	89	0.67*	145	0.81***	0.11	497	0.06
	(0.31)	(99)	(0.36)	(181)	(0.28)	(0.22)	(305)	(0.11)
Control mean	3.74	423	4.90	878	2.86	1.72	1,242	0.43
Prior buyers	−0.08	48			2.91	1.73	959	−0.63
	(2.47)	(1,028)			(2.34)	(1.23)	(1,936)	(0.95)
Control mean	14.54	3,356			11.53	2.51	6,075	2.00
Returning non-buyers	0.11	113**	0.40*	214**	0.39**	−0.02	371*	0.08
	(0.20)	(57)	(0.22)	(101)	(0.18)	(0.13)	(194)	(0.07)
Control mean	3.24	297	3.96	569	2.48	1.32	993	0.32
p: equal effects, four classes	0.081	0.570			0.249		0.459	
p: new = returning non-buyers	0.897	0.509			0.705		0.151	
Panel B. *All brokers (robustness)*								
New customers	0.10	53	0.18	52	0.27**	−0.08	12	−0.01
	(0.14)	(44)	(0.16)	(65)	(0.11)	(0.09)	(121)	(0.04)
Control mean	3.11	287	3.46	452	1.98	1.36	803	0.26
Returning non-buyers	0.00	133***	0.27	248***	0.40***	−0.06	420**	0.11*
	(0.17)	(49)	(0.19)	(86)	(0.15)	(0.11)	(172)	(0.06)
Control mean	3.33	300	4.07	585	2.46	1.29	992	0.31
p: equal effects, four classes	0.300	0.224			0.247		0.147	
p: new = returning non-buyers	0.788	0.102			0.644		0.053	
Broker, first-chat-day, rating FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Customers who chatted at least once in the experiment. Panel A, non-pilot brokers: 70,819 new customers, 27,942 prior visitors, 23,637 prior chatters, 1,949 prior buyers (51,579 returning non-buyers). Panel B, all brokers: 94,032 new customers, 68,383 returning non-buyers. Separate regressions by funnel stage. OLS with broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. 30-day columns: new customers, prior chatters and returning non-buyers only. Premium in yuan, winsorized at the 99th percentile of positive values. Benjamini-Hochberg q-values: Appendix Table A9. Equality tests: cluster-robust Wald tests, Treat interacted with the classes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

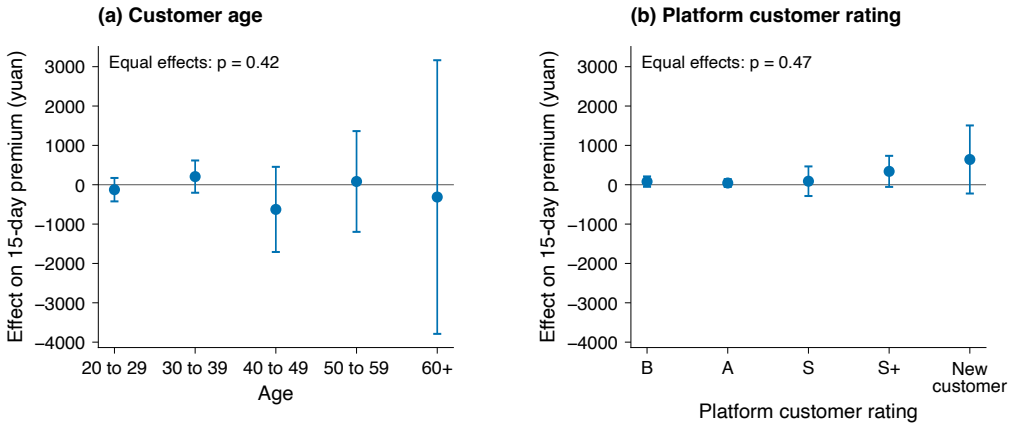


Figure A7: Effects by Age and Customer Rating

Notes: Coefficient on Treat for 15-day premium by age band (a) and by the platform's customer rating (b), customers of non-pilot brokers, with the joint p-value of equal effects across customer ratings in panel (b). 95 percent confidence intervals clustered by broker.

Table A11: Customer History, Additional Purchases, and Age

Dependent variable	15-day effect	Control mean	30-day effect	Control mean
Panel A. *Returning non-buyers*				
Premium in purchase records (yuan)	371**	1,055	415**	1,368
	(182)		(210)	
Life premium at least 50,000 yuan (pp)	0.21**	0.56	0.28***	0.70
	(0.09)		(0.10)	
Second purchase day (pp)	0.16**	0.48	0.22**	0.73
	(0.08)		(0.10)	
Panel B. *New customers*				
Premium in purchase records (yuan)	22	818	40	927
	(133)		(140)	
Life premium at least 50,000 yuan (pp)	0.03	0.45	0.03	0.52
	(0.07)		(0.07)	
Second purchase day (pp)	0.04	0.36	0.03	0.49
	(0.06)		(0.07)	
Panel C. *Age bands: joint test of equal premium effects*				
Joint p-value	0.647		0.686	

Notes: Product-coded purchase records, non-pilot brokers. Panel A has 51,579 customers and Panel B 70,819. Second purchase day counts a purchase on a date after the first recorded purchase date. Life premium thresholds refer to an individual purchase record. Panel C tests equality of premium effects across age bands. OLS with broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

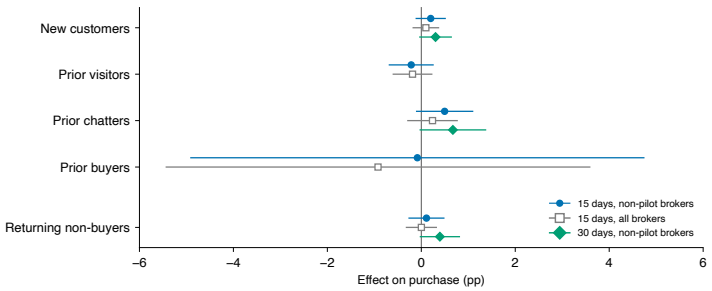


Figure A8: Purchase Effects by Prior Funnel Stage

Notes: Coefficient on Treat for 15-day purchase by prior funnel stage, non-pilot brokers (filled) and all brokers (hollow); diamonds give the 30-day purchase effect for new customers, prior chatters and returning non-buyers. 95 percent confidence intervals clustered by broker.

Table A12: Heterogeneity by Customer Rating and Age

Dependent variable:	Purchase, 15 days (pp) (1)	Premium, 15 days (yuan) (2)	Premium, first-chat day (yuan) (3)	log(1 + rounds) (4)
Panel A. *Customer rating*				
A	0.12	44	76*	0.020**
	(0.18)	(51)	(40)	(0.008)
B	0.29*	79	85	0.005
	(0.18)	(67)	(62)	(0.009)
S	−0.43	89	141	0.025
	(0.78)	(192)	(138)	(0.027)
S+	−0.15	340*	283	0.007
	(0.90)	(201)	(188)	(0.027)
New customer	−0.07	642	776***	0.024
	(1.14)	(441)	(274)	(0.037)
p: equal effects	0.653	0.472	0.252	0.792
Panel B. *Age*				
20 to 29	−1.46*	−125	−137	0.002
	(0.87)	(152)	(141)	(0.032)
30 to 39	0.17	206	199	0.010
	(0.54)	(209)	(183)	(0.020)
40 to 49	−0.52	−627	−145	−0.015
	(0.98)	(552)	(360)	(0.032)
50 to 59	1.94	83	519	−0.085
	(1.65)	(653)	(621)	(0.060)
60 or older	−3.78	−313	−403	0.012
	(3.16)	(1,774)	(345)	(0.145)

Notes: Customers of non-pilot brokers who chatted at least once in the experiment. Each cell is the coefficient on Treat in the subgroup, from a model with Treat interacted with the subgroup indicators and broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. The equality test in Panel A is a cluster-robust Wald test of the rating interactions. Premium in yuan, winsorized at the 99th percentile of positive values. Poisson pseudo-maximum likelihood with the same fixed effects and subgroup main effects gives the same conclusion: no rating or age difference is significant (equal-effects p of 0.85 for rating and 0.29 for age bands at 15 days). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A13: Purchase Baskets and the Origin of Induced Life Buyers

Panel A. *Basket shares, days 0 to 14 (pp of customers)*			
Dependent variable	All customers	New customers	Returning non-buyers
No purchase	−0.30	−0.33	−0.25
	(0.15)	(0.19)	(0.23)
Control share	96.35	96.67	95.93
Health only	−0.03	−0.02	−0.06
	(0.08)	(0.11)	(0.12)
Control share	1.21	1.19	1.23
Life without health	0.35	0.36	0.35
	(0.11)	(0.13)	(0.18)
Control share	2.03	1.81	2.34
Combined	−0.02	−0.01	−0.05
	(0.05)	(0.05)	(0.08)
Control share	0.40	0.33	0.50

Panel B. *Share of induced life buyers drawn from each basket, all customers*			
Assumption	From no purchase	From health only	Induced life mass (pp)
None	[0, 1]	[0, 1]	[0.35, 2.39]
Life buyers never leave life	[0, 1]	[0, 1]	[0.35, 0.73]
and no purchase is lost	[0, 0.83]	[0, 1]	[0.35, 0.73]
Switches only into life or combined	[0, 0.83]	[0, 0.10]	[0.35, 0.73]
Switches only into life (point)	0.83	0.10	0.35
	[0.34, 1]	[0, 0.49]	[0.14, 0.57]

Notes: Panel A: coefficient on Treat for mutually exclusive baskets from purchase records on days 0 to 14 (life without health may include other lines; combined = life and health, or other lines only), with broker bootstrap standard errors in parentheses; ability-index sample of non-pilot brokers (118,639 customers; 68,397 new customers; 50,242 returning non-buyers). Panel B: linear-programming bounds on the transitions between Control and Treat baskets consistent with the two groups' basket shares, under each assumption; brackets give the identified set at the estimates, and in the last row the 95 percent Imbens and Manski band. Random assignment identifies each group's basket shares, not individual transitions; only the last row is point identified, by assumption. In more than half of bootstrap draws the estimated shares violate the directional assumptions of the last two rows and are projected to the nearest consistent shares.

Table A14: Health Insurance Is Not Displaced

		Health		Life	
	Customers	Effect	Control (%)	Effect	Control (%)
Panel A. *Purchase on days 0 to 14, by predicted health propensity*					
Quintile 1 (lowest)	24,863	−0.10	0.80	0.59***	1.44
		(0.14)		(0.20)	
Quintile 2	20,643	0.14	0.67	−0.08	1.93
		(0.15)		(0.25)	
Quintile 3	25,554	−0.14	1.35	0.29	2.17
		(0.18)		(0.23)	
Quintile 4	28,378	0.01	1.21	−0.03	2.77
		(0.19)		(0.27)	
Quintile 5 (highest)	24,909	0.09	2.42	0.82***	3.58
		(0.25)		(0.30)	
Top decile	12,436	0.18	3.12	1.38***	4.47
		(0.43)		(0.47)	
Panel B. *Purchase on days 15 to 45*					
All customers	124,347	0.08***	0.18	0.17**	0.74
		(0.03)		(0.07)	
Returning non-buyers	51,579	0.16***	0.20	0.29***	0.88
		(0.05)		(0.11)	

Notes: Customers of non-pilot brokers. Coefficient on Treat in percentage points. Panel A: customers grouped by their predicted probability of buying health insurance within 15 days, from a logit fitted on Control customers only with pre-assignment characteristics (customer rating, prior funnel stage, prior visit days and chat rounds with the broker, first-chat weekday), with five-fold cross-fitting; out-of-fold area under the curve 0.62. If access diverted customers from health to life insurance, the health effect would be negative where health purchases are most likely. Panel B: any purchase of the product on days 15 to 45 after the first chat, from purchase records. Broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. $^{***}\ p < 0.01$, $^{**}\ p < 0.05$, $^{*}\ p < 0.10$.

Table A15: Cumulative Purchases by Product and Day since the First Chat

Purchase by day	0	1	3	7	14	29
All customers						
Life	0.16**	0.17**	0.25***	0.26**	0.33***	0.47***
	(0.07)	(0.08)	(0.09)	(0.10)	(0.11)	(0.12)
Control mean (%)	1.07	1.43	1.70	2.03	2.36	2.69
Savings-type life	0.12***	0.15***	0.20***	0.20***	0.26***	0.30***
	(0.04)	(0.05)	(0.06)	(0.07)	(0.08)	(0.08)
Control mean (%)	0.35	0.53	0.69	0.91	1.10	1.31
Health	−0.02	−0.02	−0.05	−0.03	−0.03	0.00
	(0.06)	(0.06)	(0.07)	(0.08)	(0.08)	(0.09)
Control mean (%)	0.66	0.79	0.94	1.12	1.31	1.47
Returning non-buyers						
Life	0.16	0.17	0.30**	0.26	0.39**	0.58***
	(0.11)	(0.13)	(0.15)	(0.17)	(0.18)	(0.20)
Control mean (%)	0.95	1.36	1.66	2.12	2.48	2.90
Savings-type life	0.13*	0.19**	0.30***	0.31***	0.43***	0.51***
	(0.07)	(0.08)	(0.10)	(0.12)	(0.12)	(0.14)
Control mean (%)	0.36	0.55	0.72	1.01	1.19	1.47
Health	0.01	0.01	−0.04	−0.01	−0.02	0.03
	(0.08)	(0.09)	(0.10)	(0.11)	(0.13)	(0.14)
Control mean (%)	0.53	0.66	0.85	1.07	1.32	1.50
Share of buyers by day 29 who had bought by day k (%), Control / Treat						
Life	40 / 39	53 / 50	63 / 62	76 / 73	88 / 85	100 / 100
Savings-type life	26 / 28	40 / 41	52 / 54	69 / 68	84 / 84	100 / 100
Health	45 / 45	54 / 54	64 / 63	76 / 76	90 / 88	100 / 100

Notes: Customers of non-pilot brokers (124,347) and returning non-buyers among them. Coefficient on Treat, in percentage points, for an indicator of at least one purchase of the product from the first-chat day (day 0) through day k, from purchase records (one per customer, day and product). Savings-type life: life policies of 5,000 yuan or more in round thousands. Health insurance combines critical-illness and medical insurance. Broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Under weak monotonicity restrictions (access never moves a life buyer out of life, or never stops a purchase), the share of induced life buyers drawn from health can be anywhere between 0 and 1. Under the directional assumption the 95 percent band for the no-purchase share runs from 0.34 to 1.

Buyer-level comparisons. The life share of purchase records is 3.6 percentage points higher among Treat buyers, while neither the number of records per buyer nor premium per record changes significantly. Median premium per buyer is 1,602 yuan in the Treat group and 1,082 yuan in the Control group; the 75th percentile is 20,000 against 11,800 yuan. Because access can change who buys, these comparisons describe buyers rather than how a given buyer's purchases changed.

Figure A9 describes premium among buyers; under Lee bounds, the effect on premium per buyer in the main sample lies between 803 and 1,899 yuan.

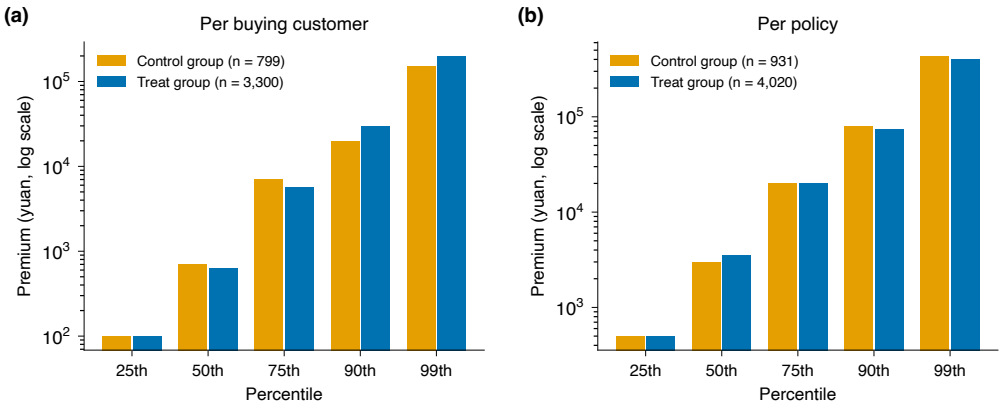


Figure A9: Premium among Buyers

Notes: Quantiles by group, customers of non-pilot brokers. Panel (a): the platform's 15-day premium field among customers with a purchase in that window. Panel (b): positive premiums in product-coded purchase records over the same window. Descriptive: the tool may change who buys.

The 30-day outcome in two parts. On the ability-index sample, the first fortnight (the platform field) gives a premium effect of 93 yuan (standard error 41) and the ability interaction of Section 4. The second fortnight (purchase records) adds purchases, 0.155 percentage points (0.065) for all customers and 0.263 points (0.106) for returning non-buyers, and premium of 62 yuan (44) and 91 yuan (92), not statistically significant; its ability interactions (39 and 53 yuan) are not statistically significant either. The ability gradient in 30-day premium therefore comes from the first fortnight.

B.4 Timing, Persistence and Later Contact

Conversation hazards. The daily model uses the 1,865,205 customer-days of days 0 to 14 for the 124,347 main-sample customers, from purchase and chat records. Customers

are active on 3.08 percent of days 1 to 14; 4.02 percent buy within days 0 to 14 and 2.00 percent on day 0. Linear-probability hazards with broker, first-chat-date, customer-rating and day fixed effects give: continuation (active on day d) +0.116 percentage points for Treat (0.036) and −0.456 (0.216) for Treat × active the previous day, where continuation for customers active the previous day is about 12 percentage points; first purchase on days 1 to 14 +0.002 (0.004); purchase on day 0 +0.175 (0.098). None of these hazards varies with broker ability, but among buyers a broker one standard deviation more able sells a life policy 4.3 percentage points more often with the tool (standard error 1.8). A Markov decomposition built on the Control hazards gives a modelled 15-day purchase effect of 0.367 percentage points, of which 0.312 when only the purchase hazards change and 0.052 when only continuation changes. In terms of the main text: the day-0 purchase effect of 0.18 points is on a base of 1.93 percent; access raises the chance that an inactive customer returns on a later day by 0.12 points, about 4 percent of a 3.1 percent base, but slightly lowers continuation among customers who chatted the day before. The purchase effect works through conversion within the conversation, not through reduced churn.

Effects by first-chat week. The 15-day premium effect is 152 yuan (standard error 61) for customers who first chatted in week 1 and 60 and 10 yuan in weeks 2 and 3; no pairwise difference survives a Holm adjustment. Compared with each broker's own pilot-period conversion and premium per customer, the decline comes from Treat customers' outcomes falling over the weeks (Figure A10). This pattern fits a fading novelty effect or later cohorts benefiting less; it does not support growing spillovers into Control conversations. The ability gradient in Section 4 does not fade over the same weeks. In the sample of all brokers every robustness estimate is similar or somewhat smaller.

Cumulative effects. In Figure 5 the cumulative premium estimates after 30 days are about 200 yuan for all customers and 431 yuan for returning non-buyers. Purchases during full rollout alone are 0.22 points higher, although brokers can then use the tool for both groups (Table A16). The cumulative premium difference is significant for returning non-buyers (460 yuan by full-rollout week 7) but not for all customers or new customers. Figure A11 plots the effects in the experiment window and in the first weeks of full rollout.

Follow-up contact. Figure A12 plots customer contact and transactions after the initial conversation. Among returning non-buyers the increase in return chat on days 15 to 45 is 1.76 percentage points. These estimates keep every randomly assigned customer in the denominator. Within the first two weeks, access raises log chat-window rounds by

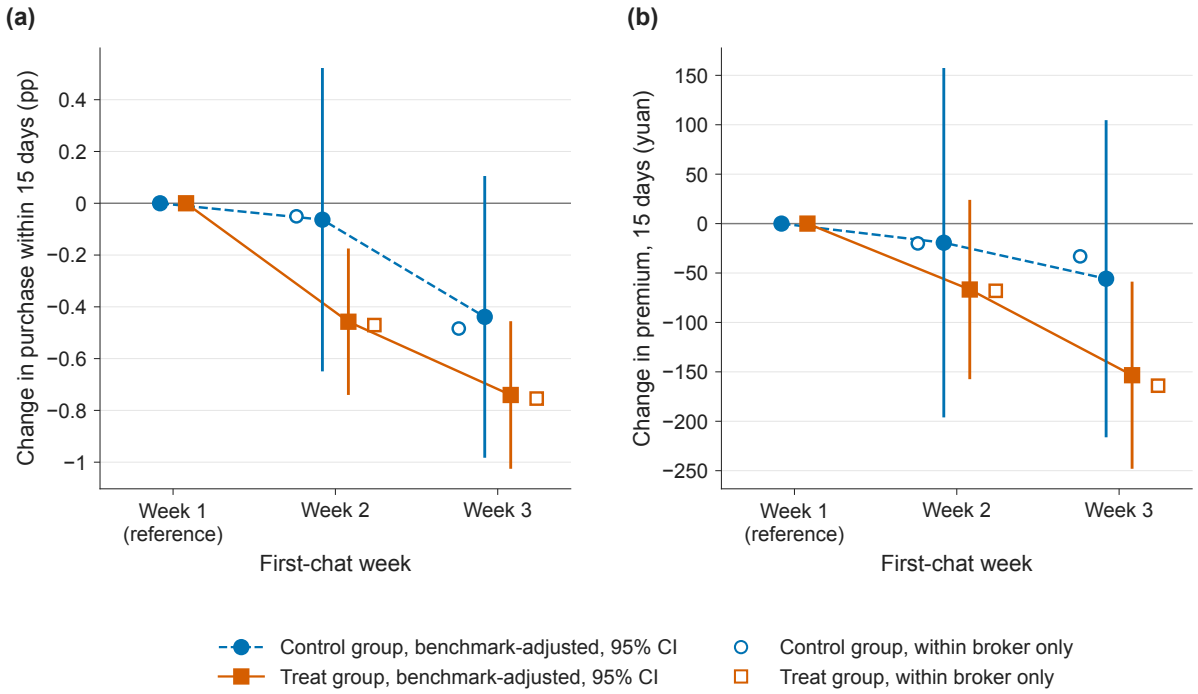


Figure A10: Which Group Moves over the Experiment Weeks

Notes: Customers of non-pilot brokers. Change in (a) purchase within 15 days and (b) 15-day premium for customers who first chatted in weeks 2 and 3 relative to week 1, separately for the Control group (circles) and the Treat group (squares). Filled markers adjust for each broker's pilot-period conversions and premium per opened customer; hollow markers use within-broker comparisons only. 95 percent confidence intervals clustered by broker. By first-chat week, the 15-day premium effect is 152 (61), 60 (77) and 10 (69) yuan, and the life-purchase effect 0.55 (0.18), 0.01 (0.21) and 0.29 (0.20) percentage points; equality across weeks is not rejected ($p = 0.27$ and 0.15).

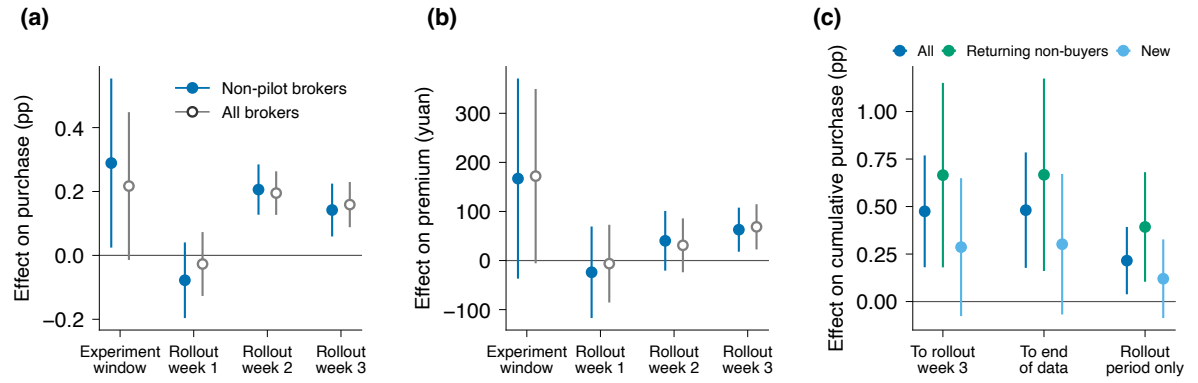


Figure A11: Persistence into Full Rollout

Notes: Panels (a) and (b): effects on purchase and premium in the experiment window and in weeks 1 to 3 of full rollout, non-pilot brokers (filled) and all brokers (hollow). Panel (c): effects on cumulative purchase to week 3 of full rollout, to the end of the data, and in the full-rollout period only, for all customers, returning non-buyers and new customers (Table A16). 95 percent confidence intervals clustered by broker.

Table A16: Initial Access to AI Advice: Cumulative Purchases and Premium through Full Rollout

	To full-rollout week 3		To full-rollout week 7		Full-rollout period only	
Dependent variable:	Purchase (pp)	Premium (yuan)	Purchase (pp)	Premium (yuan)	Purchase (pp)	Premium (yuan)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. *All customers*						
Treat	0.47***	230*	0.48***	198	0.22**	35
	(0.15)	(126)	(0.15)	(132)	(0.09)	(74)
Control mean	4.56	1,296	4.83	1,451	1.52	549
Panel B. *Returning non-buyers*						
Treat	0.67***	499**	0.67**	460**	0.39***	129
	(0.25)	(214)	(0.26)	(226)	(0.15)	(139)
Control mean	4.95	1,425	5.33	1,657	1.77	695
Panel C. *New customers*						
Treat	0.29	58	0.30	52	0.12	−13
	(0.19)	(143)	(0.19)	(146)	(0.11)	(72)
Control mean	3.81	953	3.96	1,009	1.06	331
Broker, first-chat-day, rating FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Customers of non-pilot brokers who chatted at least once in the experiment. Panels A to C: 124,347, 51,579 and 70,819 customers. OLS with broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. Outcomes cumulate purchases with any broker from the first-chat day to the end of the window (one record per customer and day); columns (5) and (6) count only purchases in full rollout, when all customers had the tool. Premium winsorized at the 99th percentile of positive values within each window. Poisson pseudo-maximum likelihood with the same fixed effects gives premium effects of 17 and 13 percent for all customers ($p = 0.09$ and 0.16) and 36 and 29 percent for returning non-buyers ($p = 0.03$ and 0.05) to full-rollout weeks 3 and 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

0.023 and returns to chat in the second week by 0.63 percentage points (Appendix B). The joint probability of buying on days 0 to 14 and returning to chat on days 15 to 45 (0.30 points) rises about as much as buying itself on the same records (0.28 points). Defined with the platform purchase field instead, the joint events keep their sign but are not statistically significant (Table A17). Later contact can reflect service requests as well as advice.

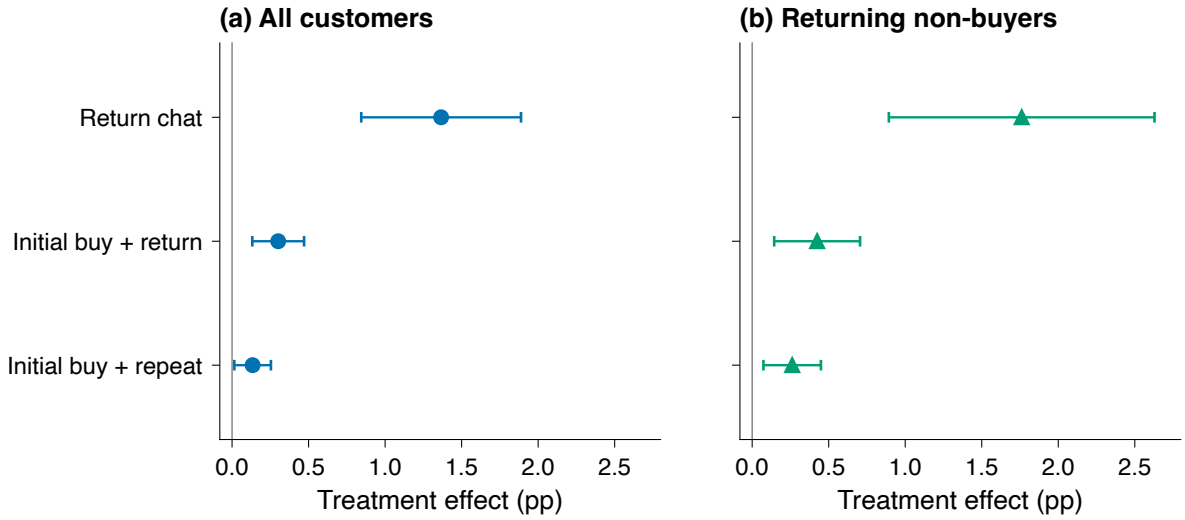


Figure A12: Customer Contact and Transactions after the Initial Conversation
Notes: All customers remain in the denominator. Return chat is on days 15 to 45. Purchase and return combines an initial purchase record on days 0 to 14 with later chat. Repeat purchase is another purchase day after that first purchase and by day 45. Effects in percentage points; 95 percent broker-clustered intervals. These are joint-event effects, not effects conditional on being a buyer.

Follow-up by cohort. Customers whose first chat falls no later than Day 9, whose day 45 falls well before the end of the data, give larger follow-up effects, all significant: return chat +1.75 percentage points, purchase and return +0.39, repeat purchase +0.22 for all customers.

Later follow-up and observation. The main follow-up uses a fixed horizon from the first chat, not from the first purchase. All experiment cohorts can be observed to day 45. Initial purchase is a daily transaction on days 0 to 14; a later purchase is on another day after the first purchase and no later than day 45. The alternative platform intersections in Table A17 are smaller and not significant. Earlier-cohort estimates retain the same direction.

Definitions. All engagement outcomes count days 0 to 14 after the first chat. *Dialogue rounds* sum the platform's daily count of rounds with the broker; *chat-window (IM) rounds* count rounds in the chat window only. *Days with a record* count days on which

Table A17: Follow-up States and Alternative Purchase Intersections

Dependent variable (pp)	All customers	Returning non-buyers
No initial purchase	-0.283**	-0.324
	(0.139)	(0.228)
Buy only	-0.054	-0.172
	(0.106)	(0.171)
Buy and return, no repeat	0.203***	0.234**
	(0.070)	(0.118)
Buy and repeat, no return	0.035	0.072
	(0.032)	(0.047)
Buy, return and repeat	0.099*	0.190**
	(0.053)	(0.083)
Joint p for the states	0.0048	0.0114
Platform purchase × later chat	0.110	0.152
	(0.076)	(0.123)
Platform purchase × repeat record	0.029	0.060
	(0.045)	(0.068)

Notes: The first five states are mutually exclusive and exhaustive; their treatment effects sum to zero. Definitions and samples are as in Table 4. The last rows intersect platform buy15 with the daily-record follow-up events. The repeat event is still defined from daily records; this is not a complete reconstruction under an independent platform purchase-date measure. The alternative intersections are not significant. State changes do not identify the behavior of marginal buyers.

the customer has any record with the broker, including visits without a message. *Return to chat* indicates a chat on days 1 to 3, 4 to 7 or 8 to 14. *Days to purchase* is the number of days from the first chat to the first purchase, among buyers. *Rounds before purchase* sum dialogue rounds from the first chat up to and including the day of the first purchase. For buyers whose purchase is recorded only in the 15-day field and not on a day with a record, the purchase day is not observed; the extended definition then counts all rounds in days 0 to 14, an upper bound.

Estimates. Table A18 reports the engagement estimates. With broker, first-chat-day and customer-rating fixed effects, for customers of non-pilot brokers: log chat-window rounds rise by 0.023 (standard error 0.007) on a Control mean of 0.91; days with a record rise by 0.028 (0.009) on a Control mean of 1.74. Returns to chat on days 1 to 3 and 4 to 7 do not change (-0.16 and 0.24 percentage points), while returns on days 8 to 14 rise by 0.63 points (0.20) on a Control rate of 8.4 percent. Among buyers, days to purchase change by 0.07 (0.13) on a Control mean of 2.1 days, log rounds before purchase by 0.015 (0.036) for buyers whose purchase day is observed, and by 0.023 (0.030) under the extended definition. None of the buyer-level estimates is significant. Because the tool may change who buys, the buyer-level comparisons are descriptive.

Table A18: Customer Engagement and the Path to Purchase

Dependent variable	Treat effect	Control mean	*N*
Panel A. *All customers: days 0 to 14*			
Log(1 + dialogue rounds)	0.013**	1.47	124,347
	(0.006)		
Log(1 + chat-window rounds)	0.023***	0.91	124,347
	(0.007)		
Days with a customer record	0.028***	1.74	124,347
	(0.009)		
Return to chat, days 1 to 3 (pp)	−0.16	18.83	124,347
	(0.27)		
Return to chat, days 4 to 7 (pp)	0.24	8.31	124,347
	(0.20)		
Return to chat, days 8 to 14 (pp)	0.63***	8.41	124,347
	(0.20)		
Purchase, days 0 to 6 (pp)	0.23*	3.25	124,347
	(0.13)		
Panel B. *Buyers only* (*descriptive*)			
Days to first purchase	0.070	2.06	4,605
	(0.133)		
Log(1 + rounds before purchase)	0.015	2.53	4,605
	(0.036)		
Log(1 + rounds), extended definition	0.023	2.37	5,976
	(0.030)		

Notes: Customers of non-pilot brokers who chatted at least once in the experiment. OLS with broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. Panel A uses all 124,347 customers. Panel B conditions on buying and is descriptive. Rounds before purchase include the purchase day; the extended definition includes all rounds in days 0 to 14 when the exact purchase day is unobserved. Log coefficients refer to log(1 + count), not raw count percentages. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

B.5 Spillovers

Daily Treat share and congestion. Within a broker and day, the excess number of Treat customers over the day's Treat share has a standard deviation of about 1.2 customers within broker and is serially uncorrelated (correlation 0.02 with the previous day). One more Treat customer of the same broker on the same day changes Control customers' 15-day premium by 6 yuan (standard error 23) and their purchase probability by 0.08 percentage points (0.07); for the other Treat customers the estimates are 12 yuan (13) and −0.01 points (0.05) (Figure A13). Of 32 same-day and lagged coefficients across outcomes and groups, one is significant at 5 percent: Control customers' life purchases after a day with more Treat customers (0.17 points, standard error 0.08). At the broker-day level, premium, conversions and chat-window time do not move with the daily Treat share; phone time rises by 26 seconds, about 1 percent of the daily mean. Scaling the same-day spillover by the average 8.4 first chats per broker-day gives an implied effect of moving a broker from no access to full access of 159 yuan in 15-day premium (standard error 107), against a customer-level effect of 84 yuan in the same sample. The band is wide, from about −50 to 370 yuan, so the customer contrast neither clearly understates nor clearly overstates the deployment effect. Table A19 reports the spillover estimates. Lags of two days are all insignificant, and dropping Day 1 gives the same picture. On busier days, Treat customers' 15-day premium is lower: per log point of the broker's same-day first chats it falls by 137 yuan (standard error 62), so doubling the day's first chats goes with about 95 yuan less; the Control slope is 23 yuan (103). The difference between the two slopes is not tested.

Table A19: Spillovers from the Random Daily Treat Share

Panel A. *Per additional Treat customer of the same broker*				
	Control customers		Treat customers	
Dependent variable	Same day	Previous day	Same day	Previous day
Premium, 15 days (yuan)	6.0	31.9	12.1	−9.9
	(23.3)	(27.4)	(12.9)	(17.0)
Purchase, 15 days (pp)	0.08	0.04	−0.01	−0.07
	(0.07)	(0.10)	(0.05)	(0.05)
Life purchase, 15 days (pp)	0.09	0.17**	−0.07*	−0.07
	(0.06)	(0.08)	(0.04)	(0.04)
Premium, 30 days (yuan)	1.0	31.5	1.0	−11.0
	(32.1)	(47.8)	(21.5)	(23.9)

Panel B. *Implied effect of full coverage*				
	Treat effect	Same-day spillover	Full coverage	p
Premium, 15 days (yuan)	84.0**	8.9	159.2	0.14
	(40.1)	(11.0)	(107.5)	
Purchase, 15 days (pp)	0.13	0.02	0.28	0.44
	(0.12)	(0.04)	(0.36)	
Life purchase, 15 days (pp)	0.23**	−0.01	0.16	0.55
	(0.11)	(0.03)	(0.27)	

Panel C. *Broker-day totals, per additional Treat customer*			
	Mean	Same day	Previous day
Premium (yuan)	12,650	176 (310)	−255 (332)
Conversions	0.62	0.004 (0.006)	0.000 (0.006)
Chat-window seconds	3,959	14.4 (17.2)	12.3 (17.7)
Phone seconds	2,174	25.7** (12.3)	22.2* (12.6)

Notes: Customers of non-pilot brokers with a first chat on Days 1 to 20 (119,386 customers, 831 brokers, 14,183 broker-days with first chats; 8.4 first chats per broker-day on average). The regressor is the number of Treat customers among the broker's first chats on the day minus the number expected at that day's Treat share (leaving out the customer's own assignment), and the same quantity on the previous day; its within-broker standard deviation is 1.2. Panels A and B: broker, first-chat-day and customer-rating fixed effects, separately by group (A) or pooled with Treat (B). Full coverage = Treat effect plus 8.4 times the same-day spillover. Panel C: 15,515 broker-days of 828 brokers, broker and date fixed effects, daily broker records. Standard errors clustered by broker in parentheses. In proportions (Poisson pseudo-maximum likelihood with the same fixed effects), one more Treat customer of the same broker that day changes Control customers' 15-day premium by 0.2 percent ($p = 0.97$) and broker-day premium by 1.1 percent ($p = 0.56$). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

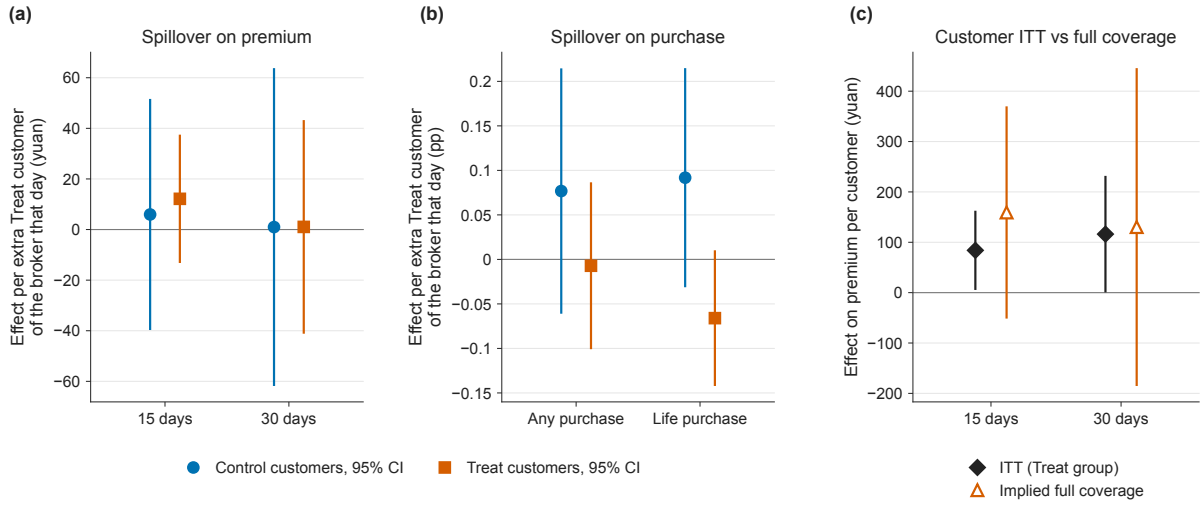


Figure A13: Spillovers from the Random Daily Treat Share

Notes: Customers of non-pilot brokers whose first chat falls on experiment Days 1 to 20 (119,386 customers, 831 brokers). Panels (a) and (b): effect of one additional Treat customer among the broker's first chats on the same day (leave-one-out excess over the day's Treat share) on Control customers (circles) and Treat customers (squares). Panel (c): the customer-level Treat effect (diamonds) and the implied effect of full coverage (triangles), the Treat effect plus 8.4 times the same-day spillover. Broker, first-chat-day and customer-rating fixed effects; 95 percent confidence intervals clustered by broker.

C Broker Ability: Additional Results

This appendix documents the ability index and the ability gradient of Section 4.

C.1 The Ability Index

Construction. The main index excludes the contemporaneous platform score and uses tenure and pilot-period daily premium, conversions and conversion rate. The premium cap used to construct the index is the positive 99th percentile of broker-day premium within the pre-experiment window (300,000 yuan). The first component explains 68 percent of the variance, with loadings 0.366, 0.502, 0.564 and 0.544. It is constructed on 820 non-pilot brokers with sufficient data. The complete non-pilot experiment sample remains 124,347 customers of 831 brokers; the common sample of the ability regressions (120,586 customers of 746 brokers) drops brokers without the inputs of every measure. Archived Table A21 is an alternative-index sensitivity result, not the main index.

What the index measures. Table A20 separates the components of the index. Entered one at a time (Panel A), tenure, daily conversions, the conversion rate and the number of customers opened per day each give a positive premium interaction of 85 to 118 yuan for all customers. Specialization in life insurance gives 2 yuan (56) for all customers and 27 yuan (90) for returning non-buyers, and its correlation with the index across brokers is 0.03. On its own it gives a life-minus-health difference in purchase gradients of 0.06 percentage points (0.17), against 0.33 for the index, while daily conversions and customers opened per day give 0.32 and 0.31. With the index in the regression (Panel B), brokers who open many customers a day with little chat-window time per day gain more than the index predicts, but only for all customers' 15-day premium (joint $p = 0.013$); for 30-day premium, purchase and returning non-buyers the added components are jointly insignificant. The index rebuilt on pilot weeks 1 to 5 and on weeks 6 to 10 correlates 0.68 across brokers; using each half as an instrument for the other (Gillen et al., 2019) gives 125 yuan (57) for all customers and 179 yuan (77) for returning non-buyers (Panel C). By tenure group (Panel D), the 15-day premium effect is 23 yuan (51) below three months and 236 yuan (98) above a year, 7 and 77 percent in proportional terms; the linear trend is 59 yuan per group ($p = 0.090$) for all customers and 100 yuan ($p = 0.050$) for returning non-buyers, while equality across groups is not rejected ($p = 0.25$ and 0.20). Brokers who survive into longer tenure may be better to begin with.

Table A20: What the Ability Gradient Measures

Dependent variable: 15-day premium (yuan)	All customers	Returning non-buyers
Panel A. *Treat × one broker component at a time (per broker SD)*		
Tenure	85** (35)	121*** (46)
Mean daily premium	55 (36)	85* (44)
Mean daily conversions	101** (39)	124** (55)
Conversion rate	118** (55)	178** (79)
Life-insurance specialization	2 (56)	27 (90)
Customers opened per day	105** (51)	94 (70)
Chat-window minutes per work day	13 (53)	51 (71)
Panel B. *Index together with the three components outside it*		
Ability index	147* (84)	204* (119)
Customers opened per day	134** (62)	39 (90)
Chat-window minutes per work day	−179** (78)	−115 (109)
Life-insurance specialization	−12 (48)	
p: three added components jointly zero	0.013	0.704
Panel C. *Measurement error in the index*		
OLS, index from pilot weeks 1 to 5	73 (55)	142* (75)
OLS, index from pilot weeks 6 to 10	99*** (33)	106** (46)
ORIV, each half instruments the other	125** (57)	179** (77)
Panel D. *Treat effect by tenure at the start of the pilot*		
Under 3 months	23 (51)	−52 (72)
3 to 6 months	75 (81)	68 (100)
6 to 12 months	25 (71)	118 (102)
Over 12 months	236** (98)	265* (145)
Linear trend per tenure bin	59* (35)	100** (51)
p: equal effects across bins	0.251	0.198

Notes: Customers of non-pilot brokers; standard errors clustered by broker in parentheses; broker, first-chat-date and customer-rating fixed effects. Components are standardized across the 820 non-pilot brokers of the index and measured over the ten pilot weeks. Life-insurance specialization is the shrunk life share of life and critical-illness conversions. Panels A and B: 120,599 customers of 750 brokers (50,247 returning non-buyers); the specialization coefficient for returning non-buyers in Panel B was not legible in the export. Panel C: the index rebuilt separately on pilot weeks 1 to 5 and 6 to 10 (broker correlation 0.68); ORIV stacks the two instrumented regressions (Gillen et al., 2019); 115,568 customers of 714 brokers. Panel D: tenure bins hold 222, 141, 277 and 186 brokers; each row is the Treat effect within the bin, all customers of non-pilot brokers. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A21: Broker Ability: Additional Outcomes and Specifications

Ability measure	Index	Index	Score	Score	Score	Tenure	Tenure
Brokers	Non-pilot	Non-pilot	Non-pilot	Non-pilot	All	Non-pilot	Non-pilot
Treat × rating controls	No	Yes	No	Yes	No	No	Yes
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. *Returning non-buyers: premium, 15 days (yuan)*							
Treat × Ability	160***	152***	152***	146**	141***	128***	116**
	(55)	(56)	(58)	(58)	(47)	(48)	(52)
Ability slope, Control group	−4		13				
	(39)		(48)				
Ability slope, Treat group	151***		176***				
	(36)		(31)				
Control mean	302	302	298	298	300	298	298
Panel B. *Returning non-buyers: purchase, 15 days (pp)*							
Treat × Ability	0.35*	0.41**	0.25	0.30	0.14	0.43**	0.49**
	(0.19)	(0.19)	(0.20)	(0.21)	(0.17)	(0.19)	(0.19)
Ability slope, Control group	0.09		0.13				
	(0.18)		(0.19)				
Ability slope, Treat group	0.49***		0.46***				
	(0.09)		(0.09)				
Control mean	3.28	3.28	3.25	3.25	3.34	3.25	3.25
MDE of Treat × Ability	0.53	0.55	0.56	0.58	0.48	0.52	0.54
Panel C. *Returning non-buyers: life policy of 100,000 yuan or more, 15 days (pp)*							
Treat × Ability	0.115**	0.095	0.065	0.050	0.080	0.085	0.052
	(0.058)	(0.062)	(0.062)	(0.062)	(0.054)	(0.053)	(0.060)
Ability slope, Control group	0.041		0.138***				
	(0.055)		(0.052)				
Ability slope, Treat group	0.131***		0.189***				
	(0.030)		(0.030)				
Control mean	0.328	0.328	0.321	0.321	0.306	0.321	0.321
Panel D. *New customers*							
Treat × Ability: premium (yuan)	54		10		−31	60	
	(48)		(46)		(44)	(51)	
Treat × Ability: purchase (pp)	−0.31*		−0.36**		−0.23	−0.24	
	(0.17)		(0.16)		(0.15)	(0.17)	
Panel E. *All customers*							
Treat × Ability: premium (yuan)	102**		68*		32	92**	
	(41)		(39)		(35)	(36)	
Treat × Ability: purchase (pp)	0.00		−0.07		−0.05	0.05	
	(0.12)		(0.13)		(0.11)	(0.12)	
Broker, first-chat-day, rating FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Comparison specification: the ability index in this archived table includes the experiment-period platform score and differs from the main pre-experiment index. Customers of non-pilot brokers (column (5): all brokers) who chatted at least once in the experiment: returning non-buyers in Panels A to C, new customers in Panel D, all customers in Panel E. Customers in Panels A to C by column: (1) 50,242, (2) 50,242, (3) 51,528, (4) 51,528, (5) 68,332, (6) 51,556, (7) 51,556. Each column: the outcome on Treat, Treat × ability and broker, first-chat-day and customer-rating fixed effects; columns (2), (4) and (7) add Treat × rating. Ability is standardized; index as in Section 4. Ability-slope rows: the outcome on ability within each group, first-chat-day and customer-rating fixed effects only. Premium in yuan, winsorized at the 99th percentile of positive values. MDE: 2.8×SE. Standard errors clustered by broker in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

C.2 The Gradient and Its Robustness

Ability quintiles. Table A22 reports, by quintile of the ability index, the Control and Treat means behind Figure 6, the premium, proportional and life-purchase effects, the effects of assignment on receipt and adoption, and the share of received suggestions that brokers send. Control premium per customer rises from 221 yuan in the bottom quintile to 418 yuan in the top for all customers and moves without a clear order for returning non-buyers; the PPML Treat effect is 18, −8, −4, 54 and 75 percent across the five quintiles. Single quintiles are noisy; the slopes in Table 5 summarize them.

Table A22: Premium, Treat Effects and AI Use by Ability Quintile

		Mean premium, 15 days		Treat effect			Effect of Treat on		Total sent
Ability quintile	Brokers	Control (yuan)	Treat (yuan)	Premium (yuan)	PPML (%)	Life purchase (pp)	Any receipt (pp)	Any adoption (pp)	share of suggestions
Panel A. *All customers*									
Bottom	177	221	247	32 (61)	18	−0.10	40.7	20.0	12.9%
Second	159	303	257	−23 (91)	−8	0.65***	42.7	18.8	10.6%
Third	147	334	303	−18 (87)	−4	0.52**	44.9	19.5	11.2%
Fourth	139	324	462	173** (87)	54*	0.21	47.9	19.8	9.1%
Top	128	418	665	304** (125)	75**	0.54*	50.2	18.1	7.7%
Panel B. *Returning non-buyers*									
Bottom	174	244	262	17 (83)	11	−0.23			
Second	159	197	274	97 (107)	52	0.49			
Third	147	434	269	−161 (161)	−42	0.24			
Fourth	139	362	540	188 (131)	45	0.23			
Top	128	281	735	468*** (155)	164***	1.23***			

Notes: Customers of non-pilot brokers with the ability index (120,599 customers of 750 brokers; about 24,000 per quintile in Panel A and 10,000 in Panel B). This sample needs only the index; the common sample of the ability regressions (120,586 customers of 746 brokers) also requires the inputs of every alternative measure. Quintiles use customer-weighted cut points of the index; each broker falls in one quintile. Mean premium: raw group means, winsorized at the 99th percentile of positive values. Treat effects within the quintile with broker, first-chat-date and customer-rating fixed effects; standard errors clustered by broker in parentheses. PPML: proportional effect, $100 \times (\exp(b) - 1)$. Any receipt and any adoption: effect of Treat on whether the broker received, or adopted, at least one AI suggestion for the customer (all significant at 1 percent). Total sent: suggestions sent in full or in part, directly or after edits, as a share of suggestions received, from the broker-day handling records (descriptive; no standard errors). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Robustness of the gradient. Table A23 reports the gradient with null-imposed wild cluster bootstrap p-values, on a customer composition common to all ability quintiles, without winsorizing, leaving out one broker at a time, and jointly with life-insurance specialization and pre-period conversion volume. The common composition restricts to customer cells (prior funnel stage, customer rating, week, weekend and age) present in every quintile and reweights them; it gives 83 yuan (38), and the drop comes from trimming cells without support, which removes relatively more customers of top-quintile brokers. Without winsorizing the gradient is 153 yuan (53). Tenure's 30-day premium interactions are positive but not statistically significant. Volume and the index are too collinear (correlation 0.79) to be separated; together they carry the gradient, but neither

is individually precise. Specialization explains none of it. Its measure is noisy across the two halves of the pilot (correlation 0.23 after shrinkage), so this is a weak test of a task-specific advantage. Archived estimates with an alternative index that includes the experiment-period platform score are in Table A21.

Table A23: The Ability Gradient: Customer Composition, Scale and Inference

Treat × ability index; dependent variable: 15-day premium (yuan)	All customers	Returning non-buyers
Baseline (Table 5)	107** (42)	150*** (56)
Wild cluster bootstrap p, null imposed	0.010	0.005
Common customer composition across ability quintiles		
Trimmed to cells present in every quintile	84** (36)	126** (55)
Trimmed and reweighted to a common mix	83** (38)	122** (61)
Treat × customer covariates added	110*** (40)	146** (58)
Scale and influence		
Premium not winsorized	153*** (53)	184** (79)
Leave one broker out: range of estimates	91 to 126	124 to 172
Specialization and volume, jointly with the index		
Ability index	55 (63)	95 (82)
Life-insurance specialization	7 (55)	33 (87)
Pre-period conversion volume	96 (89)	103 (103)

Notes: Customers of non-pilot brokers (120,586 customers of 746 brokers; 50,242 returning non-buyers). Broker, first-chat-date and customer-rating fixed effects; standard errors clustered by broker in parentheses. Wild cluster bootstrap: Rademacher weights by broker, 999 draws. Common composition: cells defined by prior funnel stage, customer rating, week, weekend and age band; cells without at least five customers in every ability quintile are dropped (3.0 percent of all customers and 4.9 percent of returning non-buyers). Leave one broker out: the gradient re-estimated without each broker in turn. Specialization: shrunk life share of life and critical-illness conversions in the pilot weeks; its correlation with the index across brokers is 0.02, and volume's is 0.79. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A24 reports the interactions on a proportional scale.

Table A24: Proportional Revenue Interactions with Pre-Experiment Ability

Customers	Measure	γ	SE	p	$100(e^{\gamma}-1)$	95% CI
All customers	Index	0.213	0.109	0.051	23.7	-0.1 to 53.3
	Tenure	0.222	0.122	0.068	24.8	-1.6 to 58.4
Returning non-buyers	Index	0.336	0.159	0.034	40.0	2.5 to 91.1
	Tenure	0.364	0.186	0.050	43.8	-0.1 to 107.0

Notes: Broker-FE PPML for 15-day premium with Treat, Treat × standardized ability, first-chat-date and customer-rating effects. Broker-clustered inference. Percentage columns transform the interaction coefficient: they measure the change in the treatment revenue ratio per standard deviation, not a percentage-point change in the treatment effect. The profile estimator was checked against a dummy-variable GLM, and alternative starts and optimizers yielded the same baseline solution (Appendix discussion).

C.3 Products and Skill Intensity

Products and policy size. Table A25 reports the product gradients behind Figure 7. Life-insurance purchase effects for all customers are 0.27, 0.34 and 0.43 percentage points at P25, P50 and P75 of ability; among returning non-buyers they are 0.05, 0.31 and 0.64 points, and the lowest is not statistically significant. Health-insurance effects are not statistically significant at any of these points. The gradient in buying at least 30K yuan within 15 days is 0.14 percentage points (0.07) among returning non-buyers; equality of the gradients across the three size groups has $p = 0.076$.

Table A25: Skill Gradients in Life- and Health-Insurance Purchases

Treat × ability (pp per SD)	All customers	Returning non-buyers
Life-insurance gradient	0.130	0.486***
	(0.109)	(0.168)
Health-insurance gradient	-0.204**	-0.178
	(0.092)	(0.134)
Life minus health	0.333**	0.664***
	(0.146)	(0.211)
Product-difference raw p	0.0225	0.0016
Wild cluster bootstrap p	0.022	0.002
Holm p, primary family (test M3)	0.044	
Holm p, six-test product family	0.112	0.0096

Notes: Stacked product equations use the pre-experiment index and product-specific broker, date and customer-rating effects. Broker, first-chat-date and customer-rating fixed effects; standard errors clustered by broker in parentheses. The two equations retain their joint covariance. Samples: 120,586 and 50,242 unique customers. Wild cluster bootstrap: null imposed, Rademacher weights by broker, 999 draws. The primary family is tests M1 to M3 of Section 3.1. The six-test product family uses the life-minus-health gradient, the sum of the two purchase-indicator gradients, and the equality of three window-premium-category gradients, for each population. Adjusted values are calculated from exported rounded raw p-values. Product-coded daily records are distinct from platform purchase fields. An effect on one product is not an identified switch from another product.

Table A26 reports product effects at quantiles of pre-experiment ability.

Skill-intensity cells. Table A27 reports the cell estimates behind Figure 8 and the errors-in-variables slopes across cells. All fourteen pre-tool slopes are positive and precisely estimated. The tool's interaction with ability is −0.21 percentage points in health insurance for new customers, −0.19 for returning non-buyers, and −0.24 in new customers' policies below 5K yuan. Among returning non-buyers the errors-in-variables slope across the seven cells is 2.13 (95 percent interval −0.11 to 6.54), and across the three non-overlapping size bins 2.05 (−0.04 to 7.97). The comparison is descriptive for three reasons. The cells overlap (any purchase and premium sum over the size bins). Among the cell interactions only the returning-non-buyer life gradient is significant at 5 percent. And the sign depends on scale: across life and health, health has the larger

Table A26: Product Effects at Pre-Experiment Ability Quantiles

Customers	Product	P25	P50	P75
All customers	Life	0.273**	0.342***	0.431***
		(0.124)	(0.109)	(0.131)
	Health	0.123	0.014	-0.127
		(0.091)	(0.081)	(0.109)
Returning non-buyers	Life	0.049	0.307*	0.640***
		(0.209)	(0.183)	(0.210)
	Health	0.096	0.001	-0.121
		(0.135)	(0.126)	(0.166)

Notes: Effects in percentage points, with broker-clustered standard errors. Linear combinations of Treat and Treat × ability in stacked OLS. Customer-weighted ability quantiles are fixed at their observed values; they are not independently estimated subgroup regressions. Health effects are not significant at these points. A non-significant effect does not establish zero effect.

pre-tool slope relative to its mean yet the tool flattens it (slope -2.59), and in percentage points rather than relative to cell means the relation across size bins is negative (-1.07, interval -2.45 to 0.17). The Control-group ability slope measured in the experiment, without broker fixed effects, has standard errors three to four times those of the pilot-period slope and is mechanically negatively correlated with Treat × ability; it is therefore not used as the measure of skill intensity.

C.4 Engagement and Conversation History

Engagement and conversation history. The Treat × ability interaction on chat-window rounds is not statistically significant. The premium gradient is 109 yuan inside the sample with chat-window contact on day 0 and 108 yuan with realized engagement controls; both checks are descriptive, because engagement responds to treatment. Tenure alone gives a positive, marginally significant interaction on chat-window contact, so the negative engagement gradient is specific to the performance components of the index. The weekly gradients (129, 57 and 126 yuan) have $p = 0.21$ for equality. The purchase effect rises by 0.45 percentage points (0.21) per standard deviation of prior chat days with the same broker and not with prior chat days with other brokers (-0.17, not significant); adding Treat × ability leaves it unchanged. Premium effects do not vary significantly with either kind of history, and effects by bins of total prior chat days are not statistically different ($p = 0.20$). The data do not show whether the tool sees history with other brokers.

Table A27: Skill Intensity by Task: Ability Slopes without the Tool and Treat × Ability

Customers	Dependent variable (15 days)	Ability slope, pilot weeks 6 to 10	Treat × ability	Treat effect	Mean of the outcome: Pilot weeks	Mean of the outcome: Control
New customers	Any purchase (pp)	0.39 (0.04)	−0.22 (0.18)	0.26 (0.17)	2.57	3.02
	Premium (yuan)	20.7 (5.1)	34.6 (55.7)	102.6 (46.0)	108	239
	Life purchase (pp)	0.15 (0.03)	−0.09 (0.15)	0.34 (0.13)	1.26	1.98
	Health purchase (pp)	0.31 (0.04)	−0.21 (0.12)	−0.01 (0.11)	1.33	1.32
	Purchase below 5K (pp)	0.26 (0.03)	−0.24 (0.15)	0.23 (0.13)	1.85	1.86
	Purchase 5K to 30K (pp)	0.06 (0.01)	0.02 (0.08)	−0.08 (0.07)	0.27	0.48
	Purchase 30K and above (pp)	0.02 (0.01)	0.06 (0.04)	0.12 (0.05)	0.10	0.21
Returning non-buyers	Any purchase (pp)	0.25 (0.06)	0.23 (0.19)	0.21 (0.20)	2.06	3.23
	Premium (yuan)	54.7 (11.4)	142.1 (74.7)	116.1 (58.4)	150	297
	Life purchase (pp)	0.13 (0.04)	0.46 (0.18)	0.38 (0.19)	1.14	2.54
	Health purchase (pp)	0.13 (0.04)	−0.19 (0.14)	−0.02 (0.13)	0.72	1.38
	Purchase below 5K (pp)	0.12 (0.04)	−0.04 (0.16)	0.00 (0.16)	1.38	1.90
	Purchase 5K to 30K (pp)	0.08 (0.02)	0.13 (0.10)	−0.01 (0.09)	0.29	0.66
	Purchase 30K and above (pp)	0.05 (0.02)	0.12 (0.08)	0.13 (0.07)	0.16	0.30

Slope of Treat × ability on the pilot-period slope (York regression, 95% bootstrap interval); relative to cell means unless stated

Returning non-buyers, seven cells	2.13 [−0.11, 6.54]; correlation 0.80
New customers, seven cells	−0.26 [−5.31, 3.51]; correlation 0.18
Three policy-size bins, both classes	2.05 [−0.04, 7.97]
Life and health, both classes	−2.59 [−12.02, −0.21]
Three policy-size bins, in percentage points	−1.07 [−2.45, 0.17]

Notes: Ability is the index rebuilt from pilot weeks 1 to 5 only (795 brokers; correlation 0.91 with the main index), so the pilot-period slope does not reuse the data that define it. Pilot-period slope: customers of non-pilot brokers whose first chat falls in pilot weeks 6 to 10, before the experiment and without the tool for these brokers (284,850 customers of 773 brokers), with first-chat-date fixed effects. Treat × ability and Treat effect: the experiment sample (115,588 customers of 717 non-pilot brokers), broker, first-chat-date and customer-rating fixed effects. Standard errors clustered by broker in parentheses. Size bins classify the platform 15-day premium of a purchase (K = 1,000 yuan). The cells overlap (any purchase and premium sum over the size bins), so the cross-cell slopes and correlations are descriptive, not tests. Among the Treat × ability coefficients, only the returning-non-buyer life gradient is significant at 5 percent. The York regression allows for sampling error on both axes; the sign of the cross-cell relation depends on the scale (relative to cell means or in percentage points).

D Adoption: Additional Results

This appendix reports further adoption results for Section 5.

D.1 Receipt and Adoption States

Receipt. Nonreceipt because a conversation ends quickly does not itself invalidate assignment as an instrument for receipt, but the binary receipt measure groups together different numbers and timings of suggestions. The receipt-scaled effects are customer-level scalings, not returns to one more suggestion.

Table A28 reports suggestion states, exposure, adoption and value per adoption by broker rank and by week of the first chat. Figure 10(a) in the main text plots adoption by rank.

Table A28: Adoption and Value per Adoption by Rank and Week

	Full adoption (%) (1)	Not adopted (%) (2)	With a suggestion (%) (3)	With an adoption (%) (4)	Premium, ITT (yuan) (5)	Normalized effect (yuan) (6)
Panel A. *Broker rank*						
L1	3.81	81.0	40.7	22.5	−94	−420
					(84)	(366)
L2	3.01	85.2	45.2	20.5	55	274
					(53)	(263)
L3	2.41	87.8	46.5	17.7	174**	1,006**
					(84)	(492)
L4	2.34	87.8	50.1	17.7	192	1,135
					(170)	(1,021)
L5	2.02	89.0	52.1	18.4	390*	2,309*
					(220)	(1,238)
Panel B. *Week of the first chat*						
Week 1 (with transition days)	2.58	87.3	46.8	17.6	149**	886**
					(60)	(358)
Week 2	2.92	85.6	45.3	19.0	56	297
					(77)	(408)
Week 3	3.08	84.3	44.6	23.9	34	140
					(69)	(286)

Notes: Archived comparison: ability-index groupings in this table use the alternative index including the contemporaneous platform score. They do not use the main pre-experiment index. Customers of non-pilot brokers. Columns (1) and (2) are suggestion-weighted shares of AI suggestions on the brokers' records; columns (3) and (4) are shares of Treat customers. Column (5) is the ITT on 15-day premium within the cell (broker, first-chat-day and customer-rating fixed effects, clustered by broker); column (6) divides it by the first stage for any adoption; the unit is per additional customer with a recorded adoption, not per suggestion. Adoption is a broker choice; the table is descriptive. Premium in yuan, winsorized at the 99th percentile of positive values. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

D.2 Adoption and Gains across Brokers

Split-sample evidence. Table A29 reports the half-to-half reliabilities (Panel A), the corrected correlations behind Figure 9 (Panel B) and the effects by adoption tercile (Panel C). The broker's premium gain has a half-to-half correlation of 0.03. Tenure

correlates 0.30 with the gain (interval −0.01 to 1.16), and the share sent −0.33 (−1.30 to −0.08). The life-purchase effect is also largest in the lowest adoption tercile. The implied standard deviation of true broker gains is about 220 to 240 yuan, but their variance is not significantly greater than zero.

Table A29: Do Brokers Who Adopt More Gain More? Split-Sample Evidence

Panel A. *Stability across the two halves of each broker's customers*

	Correlation, half to half
Adoption rate (adopted / received suggestions)	0.80
Share of customers with any adoption	0.78
Share of suggestions sent (handling records)	0.92
Broker Treat effect on 15-day premium	0.03

Panel B. *Corrected correlation with the broker's premium gain*

	Correlation	95% interval
Adoption rate	−0.39	[−1.63, −0.11]
Share of customers with any adoption	−0.33	[−1.42, 0.04]
Share of customers with any suggestion received	−0.04	[−0.47, 0.36]
Full adoption share	−0.22	[−0.83, 0.07]
Partial direct use share	−0.35	[−1.50, −0.08]
Share of suggestions sent	−0.33	[−1.30, −0.08]
Ability index	0.42	[0.08, 1.55]
Tenure	0.30	[−0.01, 1.16]

Panel C. *Treat effect by tercile of adoption measured in the other half*

	15-day premium (yuan)	Life purchase (pp)
Low adoption	161** (68)	0.52*** (0.18)
Middle	59 (73)	0.04 (0.20)
High adoption	15 (61)	0.35* (0.20)
High minus low	−147 (91)	−0.17 (0.27)

Notes: Customers of non-pilot brokers. Each broker's customers are split by the parity of the first-chat day, so the halves share no customers. Panels A and B: 654 brokers with at least 8 Control and 30 Treat customers in each half (1,379 broker-halves); outcomes residualized on first-chat-date and customer-rating fixed effects; broker Treat effect = Treat minus Control mean within the broker-half. Corrected correlation: the average cross-half covariance of the gain in one half with the measure in the other, divided by the square root of the product of the two signal variances (estimated by cross-half covariances); brokers weighted equally. Intervals from a broker bootstrap with 500 draws; they can exceed ± 1 because the signal variance of broker gains (an SD of about 220 to 240 yuan) is imprecisely estimated and not significantly above zero. Ability and tenure are fixed broker traits. Panel C: pooled regression with broker-half, first-chat-date and customer-rating fixed effects; terciles hold 304, 351 and 302 brokers; p of high minus low for premium 0.11. Poisson pseudo-maximum likelihood with broker-half fixed effects gives 54, 18 and 5 percent for the low, middle and high terciles in Panel C (high minus low −32 percent, $p = 0.20$). Standard errors clustered by broker in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D.3 Adoption over Time and Broker Time

Adoption by week. Table A31 reports adoption and effects by the customer's first-chat week. The within-broker increase in any adoption from week 1 to week 3 is 8.8 percentage points (significant at 1 percent), the adoption rate among customers who received a suggestion rises from 18.0 to 29.3 percent, and receipt stays at 43.8 and 43.9

Table A30: Broker Rank and the Effect of the Tool

Dependent variable:	All customers				Returning non-buyers			
	Purchase 15 days (pp) (1)	Premium 15 days (yuan) (2)	Purchase 30 days (pp) (3)	Premium 30 days (yuan) (4)	Purchase 15 days (pp) (5)	Premium 15 days (yuan) (6)	Purchase 30 days (pp) (7)	Premium 30 days (yuan) (8)
Panel A. *Treat effect within rank group, non-pilot brokers*								
Rank L1	0.08	−94	−0.33	−176	−0.47	−108	−0.83	−176
	(0.39)	(84)	(0.49)	(141)	(0.63)	(106)	(0.82)	(154)
Control mean	2.53	226	3.46	432	2.89	222	3.82	379
Rank L2	0.10	55	0.29*	163**	−0.04	70	0.26	250**
	(0.17)	(53)	(0.18)	(77)	(0.26)	(76)	(0.29)	(127)
Control mean	3.21	318	3.68	468	3.23	295	3.90	482
Rank L3	0.42	174**	0.61**	174	0.91**	192	1.34***	82
	(0.27)	(84)	(0.31)	(131)	(0.41)	(136)	(0.44)	(241)
Control mean	3.10	281	3.75	542	2.53	276	3.22	655
Ranks L4 and L5, pooled	−0.26	218	0.15	297	−0.10	310**	0.32	551
	(0.36)	(147)	(0.39)	(213)	(0.59)	(156)	(0.66)	(346)
Control mean	4.56	436	5.15	875	4.89	402	5.75	950
p: equal effects, four rank groups	0.434	0.111	0.520	0.174	0.186	0.089	0.094	0.102
Linear trend per rank step	−0.08	96*	0.08	96	0.21	137**	0.39	140
	(0.15)	(51)	(0.17)	(77)	(0.24)	(64)	(0.28)	(126)
Panel B. *Ranks L4 and L5 separately, 15 days, non-pilot brokers*								
Rank L4	−0.51	192			−0.27	287		
	(0.40)	(170)			(0.70)	(180)		
Rank L5	1.01	390*			0.99	401		
	(0.76)	(220)			(0.91)	(260)		
Broker, first-chat-day, rating FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Customers of non-pilot brokers who chatted at least once in the experiment. Customers by rank: L1 10,597, L2 71,141, L3 27,834, L4 12,174, L5 2,540. Each cell: Treat effect in the rank group, from a model with Treat interacted with rank groups and broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker in parentheses. Equality p: cluster-robust Wald test. Trend: Treat interacted with a rank index from 0 (L1) to 3 (L4 and L5). Panel B: ranks L4 and L5 separately. Premium in yuan, winsorized at the 99th percentile of positive values. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

percent, so the rise is in handling, not opportunity. I do not estimate a joint test of the adoption and effect trends. Figure A14 groups suggestions by the week of the suggestion record instead; its panel (b) is from an earlier run.

Figure A15 reports the first stage and the return to use by week of the first chat.

Broker time. Table A32 reports broker-day totals from the daily broker table on the random excess of Treat customers among the broker's first chats, and customer premium on the broker's same-day load. The excess has a standard deviation of about 1.2 customers. An extra Treat customer changes the broker's chat-window time that day by 14 seconds (standard error 17, on a mean of 3,959) and her chat-window time per opened customer by 0.6 seconds (0.7); it adds 26 seconds of phone time (12), about 1 percent of the daily mean, and does not change daily sales. On days when a broker has many first chats, Treat customers' premium is lower (−137 yuan per log point of first chats, standard error 62), while Control customers' premium shows no slope; the difference between the two slopes is not tested. If the pattern is real, the gain from the tool requires the broker's attention and shrinks when she is busy.

A higher randomly assigned Treat share among first chatters predicts slightly less IM time on the following broker-day. The estimates do not establish unchanged revenue or service quality: non-significant outcome responses are not equivalence tests. I do not interpret this lagged association as evidence of permanent time savings.

Table A31: Adoption and Treat Effects over the Experiment Weeks

Dependent variable	Week 1	Week 2	Week 3	Week 3 minus week 1 within broker
Panel A. *AI use for Treat customers, days 0 to 2 after the first chat*				
Any suggestion received	43.8%	43.5%	43.9%	0.4 pp
Any adoption	15.1%	17.5%	23.4%	8.8 pp***
Any adoption, among customers with a suggestion	34.5%	40.2%	53.2%	19.7 pp***
Adoption rate, among customers with a suggestion	18.0%	20.9%	29.3%	12.0 pp***
Share of suggestions sent (handling records)	8.8%	10.4%	12.1%	3.3 pp***
Panel B. *Treat effect by first-chat week*				p, equal
15-day premium (yuan)	152** (61)	60 (77)	10 (69)	0.27
15-day purchase (pp)	0.39** (0.19)	−0.19 (0.25)	−0.04 (0.24)	0.14
15-day life purchase (pp)	0.55*** (0.18)	0.01 (0.21)	0.29 (0.20)	0.15
Panel C. *Treat × ability index by first-chat week*				p, equal
15-day premium, all customers (yuan)	129*** (49)	57 (53)	126** (51)	0.21
15-day premium, returning non-buyers (yuan)	146** (67)	137* (81)	173*** (61)	0.88

Notes: Customers of non-pilot brokers. Week 1 includes first chats on the transition days. Panel A: Treat customers in cohorts whose first chat leaves a full three-day window inside the experiment (91,351 customers in these cohorts); receipt and adoption are counted on days 0 to 2 so that later cohorts are not mechanically truncated; the last row uses the broker-day handling records by record week (828 brokers). The last column is the change from week 1 to week 3 within broker. Panel B: 124,347 customers; Panel C: the ability sample of Table 5. Broker, first-chat-date and customer-rating fixed effects; standard errors clustered by broker in parentheses. No pairwise week difference in Panel B survives a Holm adjustment over the three pairs. Poisson pseudo-maximum likelihood with the same fixed effects gives 15-day premium effects of 45, 18 and 5 percent in weeks 1 to 3 (equal-effects $p = 0.47$) and Treat × ability interactions of 23, 11 and 41 percent. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

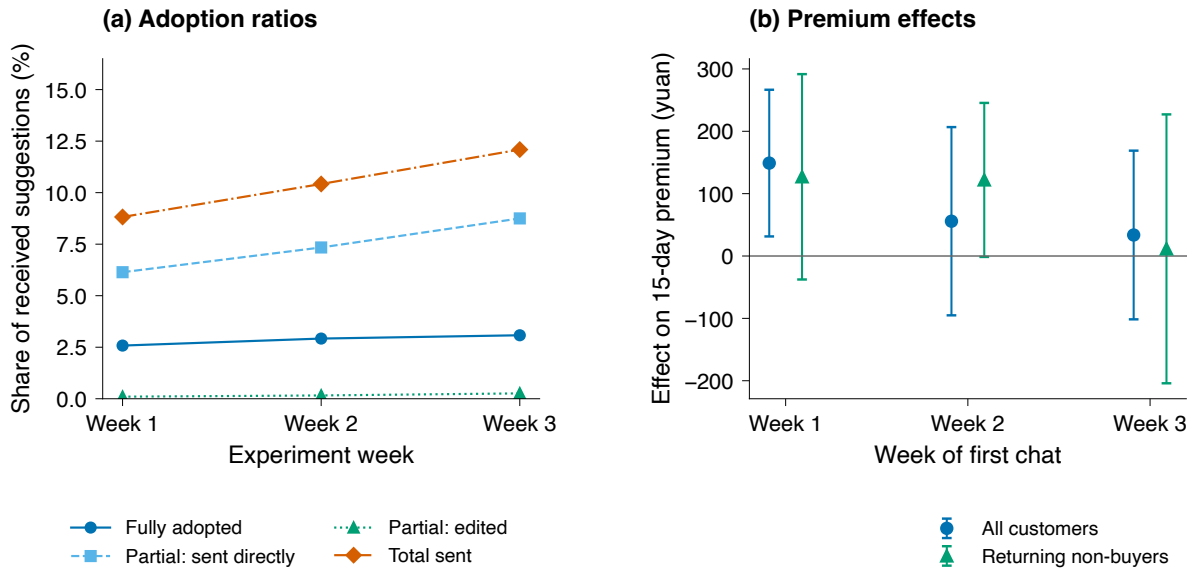


Figure A14: Recorded Adoption and Premium Effects over the Experiment

Notes: Non-pilot brokers. Panel (a) groups suggestions by record week and shows full, partial-direct, partial-edited and total-sent shares of receipt; copy-as-inspiration is excluded. Panel (b) groups customers by first-chat week and shows 15-day premium effects for all customers and returning non-buyers from an earlier run; Table A31 gives the main-specification estimates. Week 1 includes transition-day first chats. Broker, first-chat-day and rating effects; 95 percent broker-clustered intervals for the effects. Suggestion shares are descriptive.

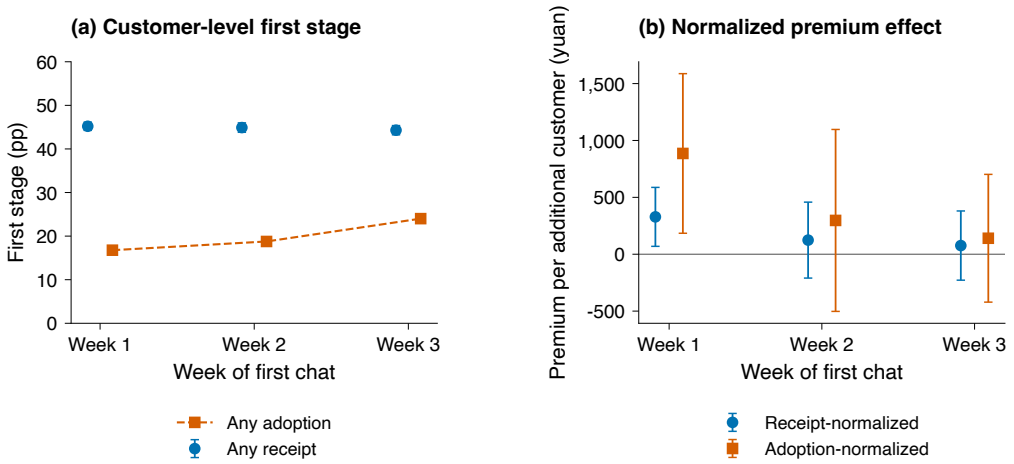


Figure A15: First Stage and Return to Use by Week of the First Chat

Notes: Panel (a): the effect of Treat on receiving at least one AI suggestion and on adopting at least one; adoption points omit confidence intervals because their standard errors are not in the archived figure data. Panel (b): the premium ITT divided by the corresponding first stage, in yuan per additional customer with receipt or adoption. The receipt normalization is an instrumental-variable effect under the assumptions in Section 5; the adoption normalization is descriptive. Week 1 includes the transition days. Customers of non-pilot brokers; displayed intervals are 95 percent and clustered by broker.

Table A32: Broker Time and Output when a Broker Serves More Treat Customers

Dependent variable	Mean	Excess Treat customers, same day	Excess Treat customers, day before	Log first chats
Panel A. *Broker-day totals*				
Chat-window seconds	3,959	14.4 (17.2)	12.3 (17.7)	1,258*** (37)
Chat-window seconds per opened customer	198	0.59 (0.71)	−1.05 (0.95)	15.7*** (2.9)
Phone seconds	2,174	25.7** (12.3)	22.2* (12.6)	407*** (29)
Customers opened	20.3	0.063* (0.038)	0.042 (0.039)	7.25*** (0.11)
Conversions	0.62	0.004 (0.006)	0.000 (0.006)	0.120*** (0.011)
Premium (yuan)	12,650	176 (310)	−255 (332)	131 (575)
Panel B. *Customer premium, 15 days (yuan), on the log of one plus the broker's first chats that day*				
Control customers				23 (103)
Treat customers				−137** (62)

Notes: Non-pilot brokers, Days 1 to 20. Panel A: 15,515 broker-days of 828 brokers from the daily broker table, with broker and date fixed effects. Log first chats: log of one plus the broker's first chats that day. Excess Treat customers: the number of Treat customers among the broker's first chats that day minus the number expected from the day's overall Treat share; because assignment is random, it is a serially uncorrelated shock with a standard deviation of about 1.2 customers. Panel B: customers of non-pilot brokers, with broker and first-chat-date fixed effects; the difference between the two slopes is not tested. In proportions (Poisson pseudo-maximum likelihood), a 1 log point increase in the broker's first chats that day lowers Treat customers' 15-day premium by 31 percent ($p = 0.016$) and leaves Control customers' premium unchanged ($p = 0.68$). Standard errors clustered by broker in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Timing and handling of suggestions. Table A33 reports when Treat customers' brokers receive suggestions, by the day's chat rounds and by day since the first chat. Table A34 reports the shares of received suggestions in each state on broker-days with at least five suggestions: their response to the random excess of Treat first chats (column 2) and their change over experiment weeks within broker (columns 3 and 4).

Table A33: When AI Suggestions Arrive

	Customer-days	Received (%)	Suggestions per round	Adoption rate (%)
Panel A. *By chat rounds on the day*				
1 round	36,297	54.4	0.71	29.9
2 to 3	16,473	70.3	0.78	18.1
4 to 5	8,474	76.6	0.81	13.4
6 to 10	9,851	82.2	0.85	11.7
11 to 20	5,564	87.3	0.90	9.9
21 or more	2,667	92.2	0.99	8.0
Panel B. *By day since the first chat*				
	Customers	Chatted (%)	Share of all suggestions (%)	Adoption rate (%)
Day 0	96,247	100.0	63.7	14.1
Day 1	89,817	10.6	8.7	9.2
Day 2	87,689	4.8	3.8	9.4
Day 3	85,486	3.6	3.3	8.3
Day 7	62,368	2.1	1.8	10.5
Day 14	32,002	1.8	2.2	12.0

Notes: Treat customers of non-pilot brokers, days 0 to 14 after the first chat. Panel A: customer-days with at least one chat round. Received: at least one AI suggestion that day; suggestions per round: suggestions divided by chat rounds; adoption rate: adopted suggestions divided by suggestions received. Panel B: customers observed on day k of their 15-day window; share of all suggestions: the day's share of all suggestions received in days 0 to 14. Descriptive.

D.4 Adoption within Conversations

Adoption within conversations. Figure A16 plots within-Treat associations between adoption intensity and outcomes; Table A35 gives alternative specifications. Among Treat customers with more than one suggestion and valid handling records, ten percentage points more adoption go with 0.36 percentage points fewer purchases and 81 yuan less 15-day premium, conditional on a linear suggestion-count control, broker, date, rating and pre-contact history; binned suggestion-count controls reduce these to 0.15 points and 48 yuan, still significant. The any-adoption indicator is not significantly associated with either outcome, and the adoption-rate interaction with ability is not statistically significant. The logs record receipt and handling, not reading, correctness or adaptation, so these are descriptive associations, not effects of adoption. The number of Treat customers excluded by the restriction to more than one suggestion is not reported in the archived output.

Table A34: How Brokers Handle AI Suggestions: Workload and Weeks

	Mean share (%) (1)	One extra Treat first chat (2)	Week 2 (3)	Week 3 (4)
Adopted (sent or copied)	14.90	0.182***	1.32***	8.61***
		(0.069)	(0.26)	(0.39)
Fully adopted	3.51	−0.018	0.39***	0.51***
		(0.033)	(0.12)	(0.13)
Partially adopted, sent directly	8.22	0.188***	0.61***	2.48***
		(0.049)	(0.18)	(0.22)
Partially adopted, edited	0.21	−0.003	0.07***	0.15***
		(0.010)	(0.03)	(0.04)
Copied	2.96	0.015	0.24***	5.46***
		(0.035)	(0.07)	(0.26)
Not adopted	84.40	−0.119*	−1.00***	−2.80***
		(0.070)	(0.26)	(0.32)
Blocked by quality check	3.63	−0.050	−0.07	−0.34***
		(0.033)	(0.10)	(0.10)

Notes: Broker-days of non-pilot brokers with at least five AI suggestions (12,250 broker-days, 810 brokers). Outcomes are shares of the day's suggestions in each state, in percent. Column (2): coefficient on the day's excess number of Treat first chats over the number expected from the day's Treat share, which is random given the broker's first chats, with broker and date fixed effects and the log of first chats. Columns (3) and (4): experiment weeks relative to week 1, with broker fixed effects and controls for the day's workload and customer mix. Standard errors clustered by broker in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

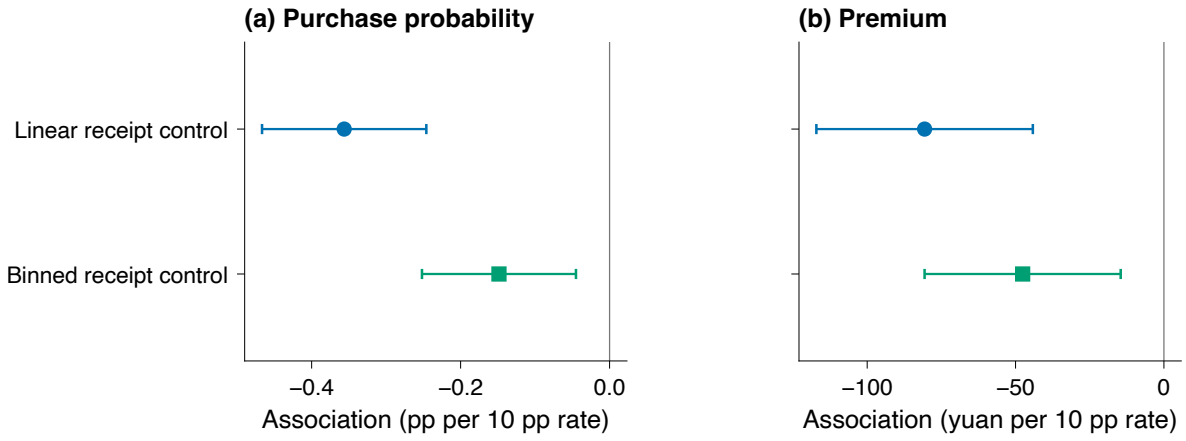


Figure A16: Adoption Intensity and Realized Outcomes

Notes: Descriptive within-Treat regressions for 31,014 customers with valid adoption and more than one received suggestion. Associations per 10-percentage-point adoption share, with either linear or binned suggestion-count controls. Broker, first-chat-date, rating and pre-contact history controls; 95 percent broker-clustered intervals. These are not randomized adoption effects.

Table A35: Adoption Intensity and Outcomes within AI-Enabled Conversations

Dependent variable	Specification	Association	SE	p
Purchase (pp)	Linear suggestion count	-0.356	0.056	0.000
	Suggestion-count bins	-0.148	0.053	0.005
	Suggestion-weighted	-0.819	0.151	0.000
	Any adoption	0.541	0.367	0.140
	Intensity among positive adopters	-0.827	0.116	0.000
Premium (yuan)	Linear suggestion count	-80.650	18.610	0.000
	Suggestion-count bins	-47.610	16.860	0.005
	Suggestion-weighted	-259.400	57.340	0.000
	Any adoption	-119.700	116.500	0.304
	Intensity among positive adopters	-122.100	32.730	0.000

Notes: Descriptive regressions among 31,014 Treat customers with more than one suggestion, valid adoption counts and the comparison ability measure. Broker, first-chat-date, customer-rating and pre-contact history controls. Rate slopes are per 10 pp, except the any-adoption dummy. Exposure is post-assignment and controlled using counts or bins; the suggestion-weighted model changes the target weighting. Positive-rate intensity is centered within positive adopters. There is no causal interpretation of changing adoption. Low suggestion counts restrict the feasible rate support; positive rates below 10 percent are unavailable for fewer than ten suggestions.

E The Pilot: Additional Results

This appendix reports further pilot results for Section 6.

E.1 Assignment, Activity and Comparison Frames

Table A36 reports the share of pilot brokers by rank on the roster at the start of the pilot.

Table A36: Pilot Assignment by Rank

	L2	L3	L4	L5	All
Brokers on the roster	715	237	97	25	1,074
Pilot brokers	118	56	19	11	204
Pilot share (%)	16.5	23.6	19.6	44.0	19.0

Notes: Brokers of ranks L2 to L5 on the roster at the start of the pilot. Pilot brokers were drawn at random within rank with rank-specific probabilities; almost no entry-rank (L1) broker was drawn. The other 49 of the 253 pilot brokers were at rank L1 or had no rank recorded on that day. Within rank, pilot and non-pilot brokers do not differ in tenure (Table 7).

Activity and comparison frames. Table A37 shows that the pilot raised the chance that a broker works at all, from the first pilot week to the end of the observation window: among the 1,074 brokers of ranks L2 to L5 on the roster at the start of the pilot (204 pilot brokers), pilot brokers are 13 percentage points more likely to work on at least three days of the first pilot week and 12 to 20 percentage points more likely to work at all in each later period. Revenue gaps on the full eligible roster therefore include an activity margin, and restricting to brokers active in the first pilot week selects on a pilot outcome; the main frame in Section 6.1 is defined before the pilot instead. Without baseline adjustment the pre-pilot frame gives slightly larger gaps than the ANCOVA (Table 8, Panel B), and the frame of brokers active in the first pilot week gives almost the same numbers. The full eligible roster gives gaps about 1.5 to 2 times as large, from 1,224 yuan in the first pilot weeks to 6,565 yuan in late full rollout. In logs, the ANCOVA premium gap is 0.10 (standard error 0.12) in the first five pilot weeks and 0.30 (standard error 0.15) in weeks 6 to 10. During the experiment pilot brokers convert 0.059 more customers per day ($p = 0.064$) and open 1.07 more ($p = 0.096$), on non-pilot means of 0.59 and 18.5.

Figure A17 compares premium and service activity per work day across phases of the rollout. Figure A18 compares raw outcomes by pilot status and group in the experiment.

Table A37: The Pilot and Whether Brokers Work at All

Dependent variable: share of eligible-roster brokers	Pilot (1)	Non-pilot (2)	Pilot effect (3)
Active on at least 3 days of pilot week 1	0.917	0.768	0.134***
			(0.024)
Any work day, pilot weeks 1 to 5	0.917	0.782	0.120***
			(0.024)
Any work day, pilot weeks 6 to 10	0.971	0.754	0.199***
			(0.019)
Any work day, experiment	0.951	0.738	0.193***
			(0.021)
Any work day, full rollout weeks 1 to 3	0.912	0.721	0.166***
			(0.025)
Any work day, full rollout later weeks	0.897	0.724	0.146***
			(0.026)

Notes: Eligible roster: the 1,074 brokers of ranks L2 to L5 on the roster at the start of the pilot (204 pilot brokers). Column (3): pilot effect with rank fixed effects; robust standard errors in parentheses. Every randomization-inference p-value from 500 within-rank permutations equals 0.002, the smallest attainable value. Within this roster, tenure at the start of the pilot does not differ (pilot minus non-pilot −0.65 months, standard error 0.49, randomization-inference $p = 0.26$). *** $p < 0.01$.

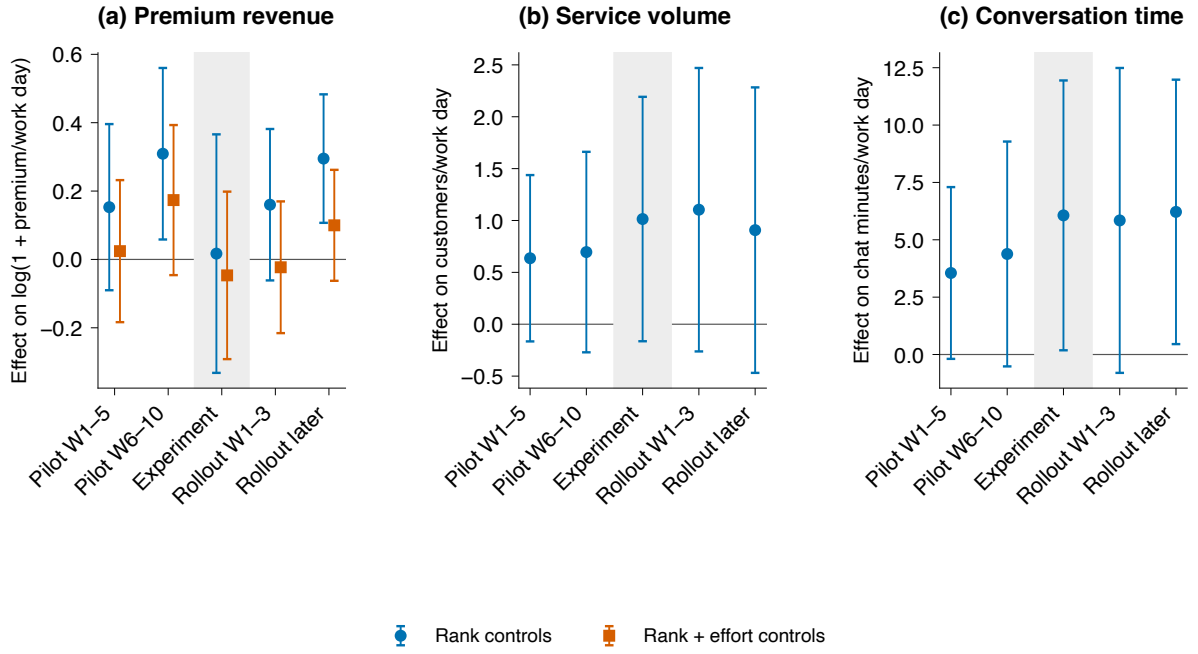


Figure A17: The Pilot at the Broker Level

Notes: Active brokers of ranks L2 to L5 (855 brokers, including 187 pilot brokers). Rank-adjusted pilot effects by rollout phase, with 95 percent confidence intervals from robust standard errors. All outcomes divide by recorded work days. Panel (a) compares log(1 + premium per work day) with and without controls for recorded effort; this adjustment is descriptive because effort responds to access. Panels (b) and (c) show customers opened and chat minutes per work day. Recorded work days do not measure total hours worked.

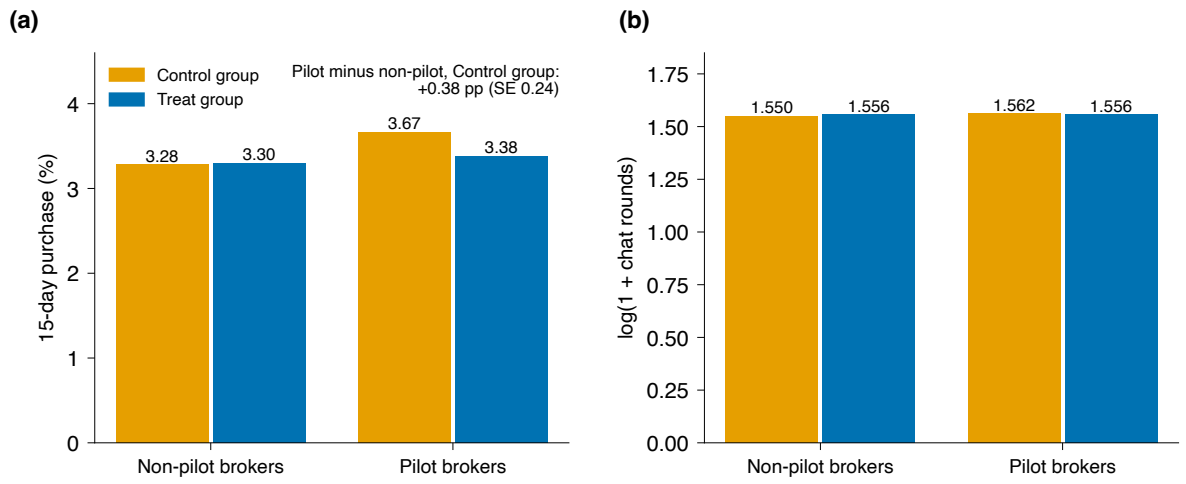


Figure A18: Outcomes by Pilot Status and Group

Notes: Raw means of 15-day purchase and log dialogue rounds in the experiment, by pilot status of the broker and group of the customer, and the pilot coefficient for Control customers. Descriptive.

E.2 Convergence and Controls

Convergence. Table A38 tests the change in the pilot gap between periods. Between the experiment and either part of full rollout, the change is small and not statistically significant in every frame and outcome: in the pre-pilot frame the later-rollout gap minus the experiment gap is 443 yuan (randomization-inference $p = 0.83$) in levels and 0.17 log points ($p = 0.35$). In yuan the gap widens from pilot weeks 6 to 10 to late full rollout (3,123 yuan, $p = 0.026$), but this widening scales with the general rise in premium per broker-day; in logs the change is 0.04 ($p = 0.84$).

Table A38: Does the Pilot Gap Shrink Once Every Broker Has the Tool?

Dependent variable:	Pre-pilot frame				Active in	Eligible
	Premium (yuan) (1)	Log premium (2)	Conversions (3)	Customers opened (4)	pilot week 1 premium (5)	roster premium (6)
Full rollout weeks 1 to 3 minus experiment	−469	−0.027	−0.005	−0.10	−522	255
	(2,280)	(0.134)	(0.034)	(0.38)	(2,270)	(2,040)
RI p	[0.834]	[0.842]	[0.886]	[0.776]	[0.788]	[0.894]
Full rollout later weeks minus experiment	443	0.174	0.008	0.00	431	550
	(2,120)	(0.146)	(0.032)	(0.53)	(2,120)	(1,900)
RI p	[0.828]	[0.351]	[0.828]	[0.994]	[0.848]	[0.737]
Full rollout later weeks minus pilot weeks 6 to 10	3,123*	0.044	0.043	0.48	3,201*	4,679***
	(1,770)	(0.209)	(0.035)	(0.67)	(1,770)	(1,610)
RI p	[0.026]	[0.836]	[0.210]	[0.497]	[0.038]	[0.004]
Brokers	862	862	862	862	858	1,074

Notes: Each entry is the change in the pilot minus non-pilot difference between two periods, from a broker by period regression with rank by period fixed effects (columns 1 to 4 also interact the pre-pilot value of the outcome with period). Outcomes per calendar day; inactive days count as zero. Frames as in Table 8. Under pure access effects the gap should fall to zero once non-pilot brokers have the tool (experiment) and once the tool is on for every customer (full rollout); the first two rows test this. Robust standard errors in parentheses; brackets: randomization-inference p-values from 500 permutations of pilot status within rank. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A39: Revenue Conditional on Recorded Effort

Dependent variable	Pilot Weeks 1 to 5 (1)	Pilot Weeks 6 to 10 (2)	Experiment Days 1 to 20 (3)	Full rollout Weeks 1 to 3 (4)	Full rollout Later weeks (5)
Panel A. *Active-broker sample*					
Log premium per day	0.175	0.404**	0.291	0.269	0.468*
	(0.127)	(0.168)	(0.234)	(0.247)	(0.252)
Log premium per day: with effort controls	0.030	0.171	−0.031	0.023	0.105
	(0.104)	(0.110)	(0.121)	(0.098)	(0.083)
Premium per day (yuan)	687	1,269**	3,990**	3,434	4,408**
	(538)	(647)	(1,860)	(2,090)	(1,810)
Premium per day (yuan): with effort controls	374	830	2,210	1,405	2,948*
	(502)	(608)	(1,730)	(1,930)	(1,790)
Conversions per day	0.016	0.040	0.074**	0.070	0.081**
	(0.018)	(0.025)	(0.036)	(0.050)	(0.040)
Conversions per day: with effort controls	−0.007	0.003	0.008	−0.002	0.012
	(0.011)	(0.014)	(0.018)	(0.025)	(0.021)
Panel B. *Eligible-roster frame*					
Log premium per day	1.121***	1.653***	1.778***	1.683***	1.721***
	(0.216)	(0.197)	(0.239)	(0.250)	(0.253)
Log premium per day: with effort controls	0.030	0.180	−0.035	0.010	0.091
	(0.104)	(0.111)	(0.117)	(0.095)	(0.079)
Premium per day (yuan)	1,224**	1,887***	6,015***	6,270***	6,565***
	(495)	(592)	(1,680)	(1,910)	(1,660)
Premium per day (yuan): with effort controls	375	702	1,752	939	2,365
	(501)	(568)	(1,610)	(1,810)	(1,670)
Conversions per day	0.066***	0.099***	0.177***	0.219***	0.196***
	(0.019)	(0.024)	(0.034)	(0.047)	(0.038)
Conversions per day: with effort controls	−0.007	0.000	0.009	0.002	0.012
	(0.011)	(0.013)	(0.018)	(0.024)	(0.019)
N with effort controls	867	854	836	813	813

Notes: Comparison frames: the active sample is selected on a pilot outcome, and the eligible roster mixes output with the pilot's effect on whether brokers work; the main frame is in Table 8. Pilot indicator and rank fixed effects. Active sample: 855 brokers before and after effort controls. Eligible roster: 1,074 in raw calendar-day regressions; the conditional rows use the reported sample sizes because per-work-day effort is unavailable for brokers with no recorded work. The raw and conditional roster rows therefore differ in sample as well as controls. Controls are work days per week, chat minutes, chat rounds, opened customers and phone minutes per work day. Conditioning on treatment responses is descriptive. Robust SE in parentheses. Later weeks include a partial eighth week. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

E.3 Customers Served and Recent Performance

Part of the pilot advantage is volume: pilot brokers open about one more customer conversation per day from the experiment on. Customers are matched to brokers by the platform, and brokers also choose how many conversations to take on, so customers opened may respond to performance. Within broker, customers opened per day rise with conversions in the previous seven days: one additional recent conversion goes with 0.66 more customers opened per day among non-pilot brokers (standard error 0.04), so a one standard deviation increase (3.2 conversions) adds about 2.1 customers, 13 percent of the mean (Table A40). Recent premium does not matter once conversions are held fixed, and the slope on conversions is the same for pilot brokers (difference 0.09, standard error 0.07). This within-broker association is not an estimate of an assignment rule, but it describes a feedback channel through which an early gain in conversions can compound into more customers served.

Table A40: Customers Opened and Recent Performance

Dependent variable: customers opened per day	Non-pilot brokers (1)	Pilot brokers (2)	Pooled: pilot × lag (3)
Conversions, previous 7 days	0.663***	0.742***	0.088
	(0.039)	(0.068)	(0.069)
Log premium, previous 7 days	−0.016	−0.051	−0.038
	(0.018)	(0.032)	(0.037)
Active days, previous 7 days	2.02***	2.58***	0.55***
	(0.08)	(0.16)	(0.17)
Mean customers opened per day	16.0	21.1	
Broker-days	98,310	23,052	
Brokers	870	204	

Notes: Eligible roster, broker by day from the start of the pilot through the first three weeks of full rollout. Dependent variable: customers opened that day. Regressors are sums over the previous seven days; broker and date fixed effects. Standard deviations of the regressors: 3.24 conversions, 4.86 log points, 2.92 active days. Column (3) is the interaction with pilot status in the pooled sample. A within-broker association, not an estimate of the platform's assignment rule. *** $p < 0.01$.

E.4 Quality and Capacity

How much of the pilot gap could come from better conversations, measured by the customer experiment, and how much from serving more customers? I scale the customer-level effect (88 yuan of 15-day premium per customer, applied to about 8.4 first chats per broker-day and to the period's premium level) to a daily flow, value the extra customers opened by pilot brokers at the non-pilot premium per customer, and treat the rest as a residual. In the first five pilot weeks, per-conversation quality from current access accounts for 242 of the 536 yuan gap (about 45 percent) and capacity for 106; in weeks 6 to 10, quality accounts for 264 of 1,076 yuan (about 24 percent) and capacity for 175,

leaving a residual of 637 yuan. From the experiment on, access no longer differs, so the quality part is zero by construction; capacity accounts for 895, 915 and 785 yuan of the gaps in the experiment and the two rollout periods, and the residual (2,737, 2,009 and 3,074 yuan) carries most of the gap. All of these parts have wide confidence bands.

The daily flow multiplies the customer-level effect by first chats per broker-day (8.4, experiment window, non-pilot brokers) times the non-pilot active-day share of the period, and keeps the proportional effect fixed across periods by scaling with the period's non-pilot premium level. Access differs by one in the pilot weeks and by zero from the experiment on. Capacity values the pilot brokers' extra customers opened at the non-pilot average rather than marginal premium per customer opened, an upper bound on the capacity part. Inputs are drawn independently (20,000 draws), which widens the residual band. Using the unscaled customer effect instead, current-access quality accounts for more than the whole gap in the first five pilot weeks (720 of 536 yuan) and 688 of 1,076 yuan in weeks 6 to 10; the post-rollout residual is unchanged because access no longer differs.

Access does not appear to free broker time. On a day when a broker happens to receive more Treat customers, chat time does not fall (14 more seconds per additional Treat customer, 95 percent interval −19 to 48 seconds) and phone time rises slightly (26 seconds). Volume differences therefore do not seem to come from time saved within conversations.

E.5 Learning, Spillover and Effort

Learning recovery. The simulation in Section 6.5 uses a broker experience stock that rises with each AI-assisted conversation and depreciates at a daily rate, affects Treat customers and possibly Control customers, and a constant selection term for pilot brokers. It is fit to 16 existing moments (pilot gaps by period, customer-level effects by week, pilot versus non-pilot Control differences, and the random dose effects) and to sets with additional candidate moments (dose lags, rollout weeks, weekly pilot gaps), over 200 simulated replications. With selection free, a slow-learning scenario that reproduces the delayed pilot gap is not recovered: the 95 percent profile sets for the depreciation rate and the learning effect cover the whole parameter grid, and the spillover of learning to Control customers is never identified. Imposing no selection recovers the parameters roughly but is rejected by the pre-pilot balance test, and in a fast-learning scenario with selection the procedure misreads fast learning as slow. Fit to the actual moments, the best-fitting stock is negative and decays within days, reproducing the decline of the customer-level effect over the experiment weeks rather than learning. The random dose within the 20-day experiment moves a broker's stock by a few customers, too little to detect learning against customer-level noise.

Dose test and Control catch-up. For each customer's first chat I count the Treat customers previously served by the broker in excess of the number expected under assignment. No coefficient is significant at the 5 percent level, and for a move from the 10th to the 90th percentile of the dose the upper confidence bounds allow effects between two and 16 times the corresponding customer-level effect (Table A41 and Figure A19). The Control-group premium difference between customers of non-pilot and pilot brokers is −145, −43 and −169 yuan in weeks 1 to 3, none statistically significant, and the purchase difference widens rather than closes (significant only in week 3, −0.92 percentage points).

Table A41 and Figure A19 report the test of learning from the random dose of Treat customers (Section 6.5).

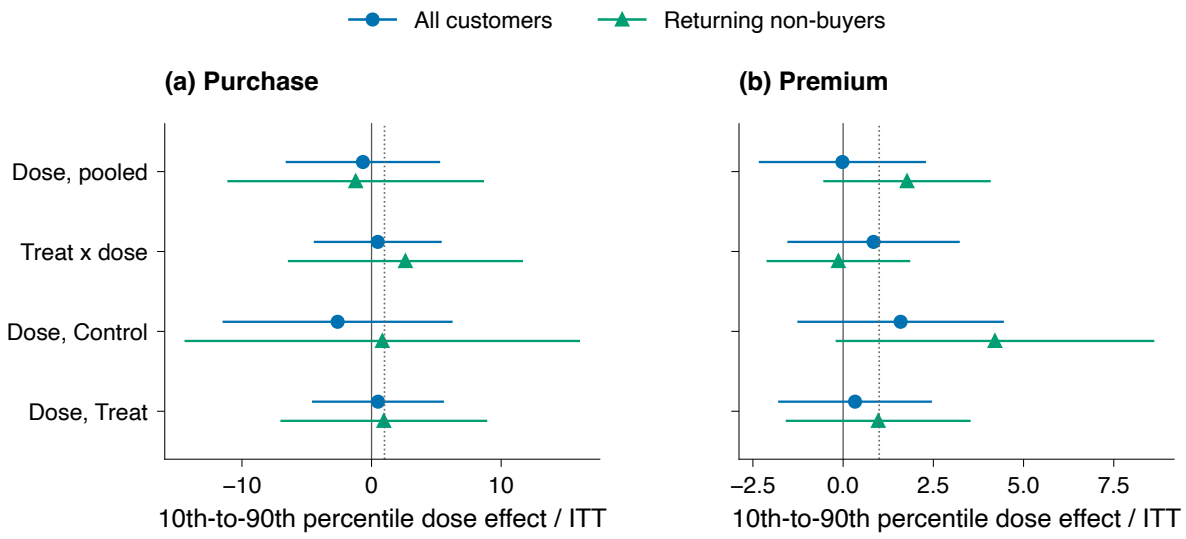


Figure A19: Bounds on Learning

Notes: For each outcome and sample of Table A41, the 95 percent confidence interval of the effect of moving from the 10th to the 90th percentile of the random dose, relative to the intent-to-treat effect of Table 2.

Effort. Table A42 reports recorded effort in the comparison frame of brokers active in the first pilot week; it does not adjust for pre-pilot effort, on which pilot brokers already differed. Pilot brokers spend 4 to 6 more minutes in chat per work day and 2 to 4 more minutes on the telephone across periods, and conditioning on recorded effort removes most of the conversion gap and a third to a half of the premium gap (Table A39). Effort is itself a possible response to access, so these conditional comparisons are descriptive. These comparison-frame tables do not adjust for pre-pilot effort; the main estimates, in the pre-pilot frame with that adjustment, are in Table A43 (Section 6.4), where the effort gaps are a few percent and not statistically significant.

Table A41: Learning from the Random Dose of Treat Customers

	Estimate per SD of dose (1)	95% CI of p10 to p90 gap effect (2)	CI upper bound / ITT (3)
Panel A. *All customers: purchase, 15 days* (*pp*)			
Dose, pooled	−0.033	[−0.770, 0.615]	6.63
	(0.151)		
Treat × dose	0.024	[−0.518, 0.629]	5.41
	(0.125)		
Dose, Control customers	−0.130	[−1.336, 0.727]	11.50
	(0.224)		
Dose, Treat customers	0.025	[−0.534, 0.650]	5.59
	(0.129)		
Panel B. *All customers: premium, 15 days* (*yuan*)			
Dose, pooled	−0.7	[−206.5, 203.0]	2.34
	(44.6)		
Treat × dose	31.8	[−136.5, 285.6]	3.23
	(45.9)		
Dose, Control customers	60.1	[−112.1, 393.7]	4.46
	(55.0)		
Dose, Treat customers	12.5	[−159.1, 217.5]	2.46
	(41.0)		
Panel C. *Returning non-buyers: purchase, 15 days* (*pp*)			
Dose, pooled	−0.057	[−1.223, 0.955]	11.13
	(0.237)		
Treat × dose	0.122	[−0.709, 1.284]	11.69
	(0.217)		
Dose, Control customers	0.039	[−1.586, 1.768]	16.10
	(0.365)		
Dose, Treat customers	0.044	[−0.773, 0.981]	8.93
	(0.191)		
Panel D. *Returning non-buyers: premium, 15 days* (*yuan*)			
Dose, pooled	85.4	[−62.1, 462.4]	4.09
	(57.1)		
Treat × dose	−6.3	[−239.8, 210.1]	2.12
	(49.0)		
Dose, Control customers	202.8*	[−23.4, 974.3]	8.62
	(109.0)		
Dose, Treat customers	46.9	[−179.6, 399.4]	3.54
	(63.0)		

Notes: Customers of non-pilot brokers who chatted at least once in the experiment. Dose: the broker's cumulative number of Treat customers before the customer's first chat minus the number expected from the same-day, same-class Treat share; standardized (raw SD 4.53; 10th to 90th percentile gap 2.34 SD). Controls: log cumulative customers and its interaction with Treat; broker, first-chat-day and customer-rating fixed effects; standard errors clustered by broker. Column (3): upper bound of the gap effect over the ITT of Table 2. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A42: Early Access and Recorded Broker Effort (Comparison Frame)

Dependent variable	Pilot Weeks 1 to 5 (1)	Pilot Weeks 6 to 10 (2)	Experiment Days 1 to 20 (3)	Full rollout Weeks 1 to 3 (4)	Full rollout Later weeks (5)
Work days per week	0.10**	0.16	0.15	0.14	0.28
	(0.05)	(0.11)	(0.14)	(0.17)	(0.18)
RI p	[0.132]	[0.204]	[0.403]	[0.397]	[0.156]
Non-pilot mean	6.79	6.48	6.25	6.10	5.91
Chat minutes per work day	3.55*	4.38*	6.06**	5.84*	6.22**
	(1.91)	(2.50)	(3.00)	(3.39)	(2.94)
RI p	[0.080]	[0.078]	[0.026]	[0.064]	[0.034]
Non-pilot mean	59.27	61.97	67.20	77.07	71.11
Phone minutes per work day	1.74	2.73**	2.91**	3.96**	1.87
	(1.11)	(1.27)	(1.43)	(1.66)	(1.37)
RI p	[0.140]	[0.036]	[0.022]	[0.006]	[0.196]
Non-pilot mean	36.44	36.33	35.24	41.13	47.48
Customers opened per work day	0.64	0.70	1.01*	1.10	0.91
	(0.41)	(0.49)	(0.60)	(0.70)	(0.70)
RI p	[0.180]	[0.164]	[0.074]	[0.116]	[0.228]
Non-pilot mean	21.72	20.79	20.65	24.70	26.92

Notes: Comparison frame: brokers of ranks L2 to L5 active on at least three days of the first pilot week, a frame selected on a pilot outcome; 855 brokers with recorded work measures, 187 pilot brokers. No adjustment for pre-pilot effort. Effects of pilot status conditional on rank. A work day has any recorded work activity; per-work-day outcomes divide by the number of such days and are not hourly measures. Parentheses: robust SE; brackets: within-rank permutation p-values (500 draws). The final period includes a partial eighth week. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A43: Early Access, Recorded Effort and Revenue per Unit of Effort

	Pilot weeks 6 to 10 (1)	Experiment Days 1 to 20 (2)	Full rollout weeks 1 to 3 (3)	Full rollout later weeks (4)	Pooled, from pilot week 6 (5)
Panel A. *Recorded effort as the outcome* (*percent*)					
Work days	0.9	0.8	0.8	3.0	1.5
	(1.6) [0.59]	(2.2) [0.74]	(2.6) [0.75]	(2.8) [0.36]	(2.1) [0.56]
Chat time	4.8	8.1**	5.5	7.5*	6.3*
	(3.3) [0.16]	(4.2) [0.08]	(4.5) [0.29]	(4.4) [0.11]	(3.5) [0.09]
Chat rounds	2.0	6.3	3.8	6.5	4.5
	(3.2) [0.60]	(4.4) [0.21]	(4.5) [0.43]	(4.7) [0.22]	(3.6) [0.27]
Telephone time	3.9	5.6	6.6	2.7	4.3
	(3.0) [0.21]	(3.8) [0.15]	(4.2) [0.10]	(3.6) [0.45]	(3.1) [0.23]
Customers opened	2.7	5.5*	4.3	4.4	4.0
	(2.3) [0.26]	(3.2) [0.08]	(3.6) [0.23]	(3.6) [0.27]	(2.9) [0.18]
Panel B. *Premium* (*percent*)					
No effort adjustment	17.3*	21.6**	11.5	13.3	15.0**
	(10.4) [0.05]	(10.3) [0.04]	(9.0) [0.18]	(8.9) [0.08]	(7.0) [0.02]
Per work day	15.9*	20.2**	9.9	10.1	13.1**
	(10.2) [0.08]	(10.1) [0.03]	(8.6) [0.22]	(8.2) [0.23]	(6.6) [0.05]
Per chat round	16.5*	16.4**	7.6	6.3	10.8**
	(9.6) [0.04]	(9.0) [0.07]	(7.4) [0.30]	(7.8) [0.43]	(5.7) [0.05]
Per customer opened	16.2*	15.9*	6.2	7.8	10.7*
	(9.9) [0.09]	(9.0) [0.08]	(7.5) [0.39]	(7.9) [0.33]	(5.9) [0.06]
Per minute of chat	12.1	13.2*	4.4	4.8	7.9
	(9.3) [0.20]	(8.4) [0.12]	(7.2) [0.56]	(7.5) [0.56]	(5.5) [0.13]
Per minute of telephone	10.5	13.3	1.3	7.7	7.6
	(9.9) [0.33]	(8.9) [0.17]	(7.5) [0.85]	(8.1) [0.34]	(5.8) [0.21]
Five effort controls	13.1	14.3*	4.0	5.2	8.2
	(9.7) [0.16]	(8.5) [0.10]	(7.3) [0.53]	(7.4) [0.47]	(5.4) [0.09]
Panel C. *Decomposition of log premium* (*log points*)					
Customers opened	0.026	0.053*	0.042	0.043	0.040
	(0.023)	(0.031)	(0.034)	(0.035)	(0.027)
Conversions per customer opened	0.028	0.032	0.011	0.036	0.028
	(0.032)	(0.030)	(0.032)	(0.029)	(0.022)
Premium per conversion	0.121	0.122*	0.051	0.045	0.076*
	(0.077)	(0.068)	(0.062)	(0.066)	(0.045)
Sum of the three parts	0.175	0.207	0.104	0.125	0.143
Log premium, no adjustment	0.159	0.195	0.108	0.125	0.140

Notes: Pre-pilot frame (862 brokers, 187 pilot brokers). Poisson pseudo-maximum likelihood on period totals per calendar day, with rank fixed effects, the pre-pilot covariates of Table 7 and the outcome's pre-pilot value; effects are $100(e^b - 1)$ percent. Panel B, rows "per: the log of the period's effort enters as an offset, so the coefficient is the premium gap less the effort gap; brokers with no recorded effort in the period are dropped (39 to 41 in column 5). "Five effort controls: the logs of the five Panel A measures as regressors. Panel C: customers opened per calendar day; conversions with the log of customers opened as offset; premium with the log of conversions as offset. Samples differ across the three parts, so their sum approximates the unadjusted log premium gap. Effort is recorded after assignment and may respond to access, so the rows in Panels B and C describe revenue per unit of recorded effort; they do not split the effect causally. Before the pilot, the same per-unit premium outcomes are 20 to 25 percent higher for pilot brokers (standard errors about 17, randomization-inference p from 0.21 to 0.31). Robust standard errors in parentheses, and stars refer to them; brackets: randomization-inference p-values from 300 permutations of pilot status within rank. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

F Model Details

This appendix gives details of Section 7 and of the relationship results in Section 4.4. Unless stated otherwise, ability in the threshold and relationship results is the index rebuilt from pilot weeks 1 to 5 with the main recipe (correlation 0.91 with the main index), with customer-weighted quintiles cut over new customers and returning non-buyers of non-pilot brokers.

Decomposition. A linear fit of the decomposition in Table 9 across the five ability quintiles attributes 23 percent of the gap to purchase probability and 77 percent to premium per buyer. On the purchase records with product categories, 51 of the 832-yuan top-minus-bottom gap is health insurance; basket composition is the only component whose interval excludes zero. By basket, access changes the share of customers buying life insurance only by -0.03 percentage points in the bottom quintile and $+1.17$ (0.45) in the top, and the share buying health insurance only by $+0.47$ (0.21) and -0.05 (0.34).

Effects across policy sizes. For each customer I sum the life premium of days 0 to 14 and record whether it reaches a threshold. Among returning non-buyers, the Treat effect relative to the pooled Control level is 17 percent for any life policy, 31 to 44 percent for thresholds from 1,000 to 30,000 yuan, 46 percent at 100,000 yuan and 68 percent at 200,000 yuan; the difference between the effects in the top and bottom ability quintiles, relative to the same Control level, is 62 percent for any life policy, 96 to 99 percent at 1,000 to 30,000 yuan and 145 percent at 100,000 yuan (0.46 percentage points, standard error 0.22). Among new customers the relative effect is about 30 percent from 1,000 to 30,000 yuan and near zero from 100,000 yuan, and the top-minus-bottom difference is small or negative at every threshold (-18 percent for any life policy, -20 percent at 200,000 yuan). In health insurance the top-minus-bottom difference is negative in both classes, -42 percent of the Control level for returning non-buyers and -25 percent for new customers, and more negative for policies of 2,000 yuan or more (-55 and -70 percent). A tool that raised every buyer's amount by the same percentage would also make the relative effect rise with the threshold; under a lognormal fitted to the Control size bins, that alone raises it by a factor of about 1.5 between 5,000 and 30,000 yuan, against 3.3 for returning non-buyers and 9 for new customers in the data, although each cell is imprecise.

Timing. In the top ability quintile of returning non-buyers the extra life purchases are spread over the first-chat day (0.60 percentage points, standard error 0.22), days 1 to 14 (0.58, 0.38) and days 15 to 29 (0.29, 0.18), and the share of life buyers who buy on the first-chat day barely differs between the groups. Access does not move the decision into

the first conversation.

History split. Returning non-buyers are split by their 70 days before the first chat: those who had chatted with the same broker (15,755 customers, about a third), those who had chatted only with other brokers (6,374) and prior visitors (26,008). Among the first, the Control rates of buying any life policy and a life policy of 30,000 yuan or more are 3.22 and 1.11 percent; access raises them by 1.04 (0.35) and 0.61 (0.21) points, and each standard deviation of ability raises these effects by 0.88 (0.33) and 0.56 (0.20) points. For customers who had chatted only with other brokers the effects and interactions are small, and for prior visitors the interaction is 0.42 (0.20) points for any life policy and 0.13 (0.15) for policies of 30,000 yuan or more; tests of equal interactions across the three splits give p-values of 0.06 to 0.08 for life insurance. Within the first split, the number of earlier chat days with the broker does not change the effect or its gradient.

Allocation of effort. For the comparison with the pilot weeks, I build the same split for customers of non-pilot brokers whose first chat inside pilot weeks 6 to 10 followed earlier contact since the start of the pilot, and stack them with the experiment's customers. Among customers who had chatted with the same broker, relative to that baseline and in units of each period's mean, the ability slope in buying a life policy of 30,000 yuan or more falls by 0.52 (0.17) for Control customers and changes by 0.00 (0.14) for Treat customers; for any life policy the changes are −0.24 (0.10) and 0.03 (0.07). Control customers of stronger brokers in this split also buy more health insurance than in the baseline (0.50 points per standard deviation, standard error 0.25). With a common 35-day history window in both periods, the Control slope still falls (−0.37, 0.18, for large life policies) but the Treat-minus-Control difference is smaller and not significant (0.27, 0.19). On days on which a broker happened to receive more Treat customers, outcomes of that broker's Control customers do not fall more at stronger brokers, so the reallocation, if any, is not day to day.

Pilot check. Pilot brokers had the tool for every customer and could not divide effort between groups. Because their pilot-period sales were already affected by the tool, I measure ability for pilot and non-pilot brokers alike by an index of the same recipe built on the days just before the pilot (correlation 0.53 with the index from pilot weeks 1 to 5) and by rank, among rank B1 to B4 brokers active before the pilot, with rank fixed effects and pre-pilot covariates. In the experiment these measures give smaller interactions than the main index among customers who had chatted with the same broker: 0.27 (0.19) and 0.15 (0.20) of the Control mean for life policies of 30,000 yuan or more, and 0.59 (0.31) and 0.12 (0.30) for premium. In pilot weeks 6 to 10, the interaction of pilot

status with the same measures, for the same customer split and in the same units, is 0.28 (0.17) and 0.35 (0.17) for large life policies and 0.04 (0.18) and −0.03 (0.22) for premium. Restricted to first chats whose 15-day window ends before the experiment, the large-life interactions fall to 0.10 and 0.23 and are not significant, and in the pilot they do not differ between customers who had chatted with the same broker and prior visitors.

Usage accounting under the main index. An accounting model splits the fitted premium effect into suggestion receipt, adoption rate and implied return per recorded adoption. Of the 388-yuan gap between the top and bottom quintiles of returning non-buyers, receipt accounts for 89 yuan, the lower adoption rate of stronger brokers for −176 yuan, and the implied return per adoption for 475 yuan. The implied return combines conversation opportunity, timing, selection and adaptation, so it is not a causal return to judgment; I treat it as secondary to the decomposition in Table 9. For a cell c defined by first-chat week, ability quintile and customer history, the model writes the assignment effect on outcome y as

$$\tau_y(c) = e(c)\,\rho_y(c), \qquad \rho_y(c) = \begin{cases} \lambda_{y,t} + \beta_y a_c, & \text{returning non-buyers,} \\ \lambda^N_{y,t}, & \text{new customers,} \end{cases} \tag{7}$$

where $e(c)$ is the assignment effect on suggestions received per customer and a_c is mean ability in the cell. Receipt reflects whether conversations last long enough for suggestions to be generated, not only broker choices. Writing $\rho_y(c) = q(c)r_y(c)$, with $q(c)$ the ratio of the adoption and receipt first stages, gives the Shapley split in Table A44. The estimation sample has 118,652 customers of 750 non-pilot brokers with the main index; the 18 cells supply 72 moments for 50 parameters. The return per suggestion rises by 43.5 yuan per standard deviation of ability (standard error 19.2), and constant returns across weeks are not rejected.

Routing customers to brokers. Cell means by customer class and broker ability quintile, with and without the tool, value routing under the observed capacity of each quintile (Table A45, Panel A). A rule that sends returning non-buyers to the strongest brokers first is worth 13.5 yuan per customer over the status quo without the tool and 86.9 yuan with it, an increase of 73 yuan (95 percent band 22 to 125). An optimal assignment computed from the cell means gives a smaller increase (53 yuan, band −6 to 80), but its value is biased upward because it maximizes over noisy cell means. The result rests on returning non-buyers: when new customers are also given an approximate ability profile, the increase from the ranked rule falls to 32 yuan and is no longer distinguishable from zero. Because part of the premium gradient among returning

Table A44: Suggestion Receipt, Adoption and Implied Return: Accounting under the Main Index

Panel A. *Premium return per AI suggestion (yuan), returning non-buyers at mean ability and new customers*				
	Week 1	Week 2	Week 3	Ability slope (per SD)
Returning non-buyers	48.5*	66.8***	68.9**	43.5**
	(26.1)	(25.7)	(33.3)	(19.2)
New customers	36.1*	39.0	2.6	
	(21.7)	(30.3)	(28.0)	
$J = 15.68$ (22 df, $p = 0.83$); no ability slope: $D = 5.94$ (2 df, $p = 0.051$); equal returns across weeks: $D = 0.73$ (4 df, $p = 0.95$)				
Panel B. *Top minus bottom quintile, fitted premium effect per returning non-buyer (yuan)*				
	Total	From receipt	From adoption rate	From return per adoption
Top (417.1) minus bottom (29.6)	387.5	88.5	−175.7	474.7
	(149.9)	(28.4)	(61.6)	(177.6)
Panel C. *Targeting access to returning non-buyers*				
Share of the fitted premium gain retained				0.61 (0.25)
Share of AI suggestions retained				0.38 (0.01)

Notes: Minimum-distance fit of $\tau_y(c) = e(c)\rho_y(c)$, with $\rho_y(c) = \lambda_{y,t} + \beta_y a_c$ for returning non-buyers and $\lambda^N_{y,t}$ for new customers, to 72 jointly estimated cell moments (premium and purchase effects, suggestions received and adoptions; 18 cells by week of first chat, ability quintile of the main index and customer history); 50 parameters. Panel B is a Shapley split of the fitted gap into suggestion receipt, adoption rate and implied return per recorded adoption; it is accounting, not a causal split, because brokers choose what to adopt and receipt depends on conversation length. Panel C restricts access to returning non-buyers holding each cell's fitted effect fixed. Broker-clustered standard errors (Panel A) and parametric bootstrap standard deviations (Panels B and C) in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

non-buyers reflects the division of effort between groups, the routing gains are upper bounds.

Full deployment. The customer-level effect holds the broker's other customers fixed. Adding the estimated same-day spillover of an additional Treat customer on the broker's other customers, times the 8.4 first chats per broker-day, gives an implied full-coverage effect of 159 yuan per customer, against a customer-level effect of 84 yuan in the same sample, or about 1,340 yuan per broker-day; the spillover term is not statistically significant, so the band is wide (−434 to 3,114 yuan per broker-day). Access does not appear to free broker time (Section E.4), so I find no support for an additional capacity gain at full saturation. The same-day spillover does not capture a broker's standing division of effort between groups, which would vanish at full deployment.

The firm's return. Premium per AI suggestion is about 40 yuan at mean ability (standard error 11.7), from 29 for new customers to 58 for returning non-buyers. Life insurance is 91 to 95 percent of premium, and industry reports put first-year commission on single-premium savings policies at about 2 to 3 percent, against about 9 to 29 percent for regular-premium savings policies, depending on the payment term, and more for long-term protection. With first-year commission rates of 3 to 35 percent, brokerage revenue per suggestion is about 1 to 20 yuan. Inference costs under assumed token counts and public list prices are 0.002 to 0.3 yuan per suggestion. Revenue exceeds inference cost by a factor of about 3 under the least favorable combination and by several

hundred under central values (Table A45, Panel C). These are ranges, not estimates: commission rates, payment terms and token use are assumptions, and development costs are excluded.

Counterfactual details. Routing uses Control means and Treat effects for five ability quintiles of returning non-buyers and one new-customer cell; Control-mean sampling error is approximated from the effect standard errors and group sizes. Capacity is brokers times mean first chats per day in each quintile; with slack, each broker can move toward her 90th-percentile day, and the congestion adjustment applies the estimated response of Treat and Control outcomes to the broker's daily load. The full-coverage effect adds the same-day spillover of one more Treat customer on the broker's other customers times mean first chats per broker-day and ignores longer lags. The firm-return calculation uses the accounting model's premium per suggestion at mean ability, first-year commission rates of 3, 20 and 35 percent, and inference costs of 2,000 to 8,000 input and 100 to 300 output tokens per call and one to three calls per suggestion at public list prices.

Table A45: Counterfactuals: Routing, Full Deployment and the Firm's Return

Panel A. *Value of routing over the status quo (yuan per customer, 15-day premium)*			
	Without the tool	With the tool	Change
Ability-ranked rule, no slack	13.5	86.9	73.4
	[−30.7, 57.4]	[64.1, 109.2]	[21.9, 125.4]
Ability-ranked rule, 25 percent slack, congestion-adjusted			95.3
			[33.2, 157.5]
Linear-program optimum, no slack	35.1	87.8	52.7
	[16.8, 86.4]	[64.8, 112.4]	[−6.1, 79.7]
Ability-ranked rule, new customers' ability profile allowed	−14.1	18.2	32.4
	[−99.5, 60.7]	[−22.1, 60.1]	[−58.4, 123.6]
Panel B. *Full deployment (15-day premium)*			
	Per customer	Per broker-day	
Customer-level effect (experiment)	88.4	744	
	[9.6, 167.1]	[81, 1,407]	
Implied effect at full coverage, with same-day spillover	159.2	1,340	
	[−51.5, 369.9]	[−434, 3,114]	
Panel C. *Firm revenue and inference cost per AI suggestion (yuan)*			
	Low	Central	High
Premium per AI suggestion	28.8	40.0	58.2
First-year commission rate (assumed)	3%	20%	35%
Brokerage revenue per AI suggestion	0.9	8.0	20.4
Inference cost per AI suggestion (assumed)	0.0018	0.028	0.28
Premium per AI suggestion that covers inference cost	0.06	0.14	0.81

Notes: Panel A: customers are assigned to ability quintiles of the main index under quintile capacity from first chats per broker-day; cell means without the tool are Control means and with the tool add the cell's Treat effect, which requires that cell means carry over to customers moved across brokers (platform matching treated as random). The ability-ranked rule sends returning non-buyers to the top quintile first without looking at cell means; the linear-program optimum maximizes over noisy cell means and is biased upward. The last row uses an approximate ability profile for new customers. 95 percent bands from 1,000 draws in brackets. Panel B: the full-coverage effect adds to the customer-level effect the same-day spillover of an extra Treat customer times 8.4 first chats per broker-day; 95 percent intervals in brackets. Panel C: premium per AI suggestion from the accounting model at mean ability (central: all customers pooled, standard error 11.7; low: new customers; high: returning non-buyers). Commission rates are assumptions anchored on industry reports, since the firm's own rates are not public; the low rate reflects single-premium savings policies (about 2 to 3 percent in the first year), which carry most of the life premium; inference cost combines assumed token counts and calls per suggestion with public list prices for large open language models. Premium is the customer's payment, revenue is gross brokerage revenue before broker pay and other costs, and development costs are excluded.