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

Yingju Ma

University of Cambridge

May 9, 2026

Preliminary draft. Please do not cite or circulate without the author's permission.

Abstract

This paper provides large-scale experimental evidence on how generative AI affects realized revenue in high-stakes financial advisory. I study a production-grade generative-AI tool that offers licensed insurance brokers real-time reply suggestions during customer advisory conversations. The field experiment, conducted at China's largest insurance brokerage Ant Insurance, covers over 1,000 brokers and premium revenue equivalent to tens of millions of US dollars. Randomization is at the customer level: brokers see AI suggestions only when serving treated customers. AI assignment raises premium revenue by 30.7 %, operating through both an extensive-margin gain in conversion (16 % relative increase) and an intensive-margin gain in per-policy premium (12 %, conditional on purchase). The setting enables a further decomposition into adoption frequency and per-adoption return: junior brokers adopt AI suggestions more frequently, whereas revenue gains are larger for senior brokers. AI exposure also induces broker learning, with gains transferring to non-AI interactions.

JEL codes: D83, G22, J24, L86, O33

Keywords: Generative AI; Productivity; Financial Advice; Field Experiment; Skill Complementarity; Human-AI Collaboration

*I am grateful to Nikhil Agarwal, Vivek Bhattacharya, Taha Choukhmane, Sara Fisher Ellison, Tobias Salz, Paul Scott, Robert Townsend, Benjamin Vatter, and Constantine Yannelis for helpful comments and discussions, and to seminar participants at the MIT IO Lunch. I acknowledge and appreciate 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. Email: ym459@cam.ac.uk.

1 Introduction

Whether generative AI raises productivity in high-stakes professional services, such as financial advisory, is among the most consequential open questions about the technology's economic impact. Deployment in such settings raises several distinctive concerns absent in lower-stakes domains. Errors are costly and often irreversible, unlike a coding task that can be re-run. Hallucinations carry direct legal and reputational liability, because the licensed professional, not the algorithm, is the agent of record. And customers may distrust visibly AI-generated communication in interactions where personal judgment is itself part of the service. Existing studies concentrate in relatively low-stake white-collar settings: customer support (Brynjolfsson et al., 2025), software development (Peng et al., 2023; Cui et al., 2026), short-form writing (Noy and Zhang, 2023), and simulated consulting tasks (Dell'Acqua et al., 2023). In each, the cognitive task is predominantly linguistic or algorithmic, and effects are measured on internal productivity proxies rather than realized market outcomes. Whether the productivity gains documented there generalize to live, regulated, advisor-mediated financial advice is unsettled.

This paper provides experimental evidence on whether generative AI raises productivity in regulated financial advisory, using a randomized field experiment at Ant Insurance, China's largest insurance brokerage. The platform serves over 600 million customers, ranks first among the 2,000+ licensed Chinese insurance intermediaries by revenue, and intermediates tens of billions of RMB in annual premium volume. More than 1,000 licensed brokers on the platform advise customers on complex and high-premium insurance products. In 2025, the platform deployed *Tongtiandai*, a production-grade generative-AI tool that offers real-time reply suggestions to licensed brokers during live customer advisory conversations. Each suggestion is delivered into the broker's chat workspace while the conversation is in progress; the broker can adopt it verbatim, edit it before sending, or discard it. The human, not the algorithm, controls what reaches the customer. The setting therefore provides a high-value, expertise-intensive, customer-facing financial-service context in which to test the productivity effect of generative AI.

The experiment ran for twenty days in mid-2025 and randomized treatment at the customer level. When a customer first contacted a broker, she was assigned to the treatment group (AI active in the broker's workspace) or the control group (AI disabled), with assignment fixed for the experiment's duration. Because each broker served customers from both groups concurrently, all comparisons are within-broker, and broker-level confounders (seniority, tenure, motivation, day-of-week effects) difference out by construction. The sample comprises roughly 1,000 licensed brokers and 165,000 customers, primarily in the life-insurance and critical-illness product lines, the plat-

form's most complex and highest-premium products, with average transaction premiums around 27,000 RMB. I observe insurance-purchase decisions over the fifteen days following first contact, capturing essentially all completed transactions.

First, on the basic question of whether AI raises productivity in a high-stakes professional service, I find a positive and economically meaningful effect. AI assignment raises premium revenue by 30.7 %. The fifteen-day conversion probability rises by 0.63 percentage points, a 16 % relative increase on a control-group conversion rate of 3.8 %, and conditional on purchase the per-buyer premium rises by approximately 12 %. The two margins reinforce each other rather than offset: AI brings more customers across the line, and the customers it brings buy larger policies.

The effect operates through actual AI use, not mere assignment. Brokers deployed the AI tool in 44.3 % of treated conversations, and the local average treatment effect on AI-deployed conversations is 1.41 percentage points, more than twice the intent-to-treat estimate. The implied compliance rate from the identity ITT $= \hat{\pi} \times$ LATE is $0.63/1.41 = 0.447$, almost exactly equal to the directly observed AI-use share of 0.443. Treated conversations in which the broker did not use AI are indistinguishable from controls; the entire ITT runs through the subset of conversations in which AI was actually deployed.

The conversion gain operates through expanded customer engagement. Cumulative dialogue rounds over the twenty-day window rise by approximately 9 %, while pre-purchase rounds conditional on purchase are essentially unchanged (-1.1 %, not significant). AI does not shorten the path to purchase for customers who were already converting; it brings additional customers, those who would otherwise have disengaged, far enough through the conversation to make a purchase decision.

The platform sorts customers into four ordered pre-experiment value tiers, B, A, S, S+ in ascending order, inferred from prior platform behavior, plus a separate "New" tier for customers with no recorded platform history. The conversion gain concentrates among lower-value tiers: the lowest tier B shows a +45 % PPML increase, tier A +26 %, and the top tier S+ is indistinguishable from zero. New customers display no detectable conversion gain but a striking +52 % increase in pre-purchase dialogue rounds, consistent with AI operating as a discovery tool for unfamiliar customers.

Second, which brokers gain more from AI? The dominant pattern in prior field experiments is leveling: low-skill workers gain more, narrowing the productivity gap. This setting inverts the pattern. Estimating treatment effects across the platform's four broker rank levels (L1 the most junior, L4+ the most senior), I find that intent-to-treat effects on the conversion rate are near zero at L1 (imprecise, given the small L1 broker count) and rise monotonically from L2 through L4+, reaching +0.90 percentage points at the most senior level; on actually-AI-deployed conversations the LATE reaches +1.88 percentage points at L4+. AI therefore *amplifies*, rather than levels, productivity dif-

ferences across brokers. The reversal is consistent with theories in which AI complements human expertise when stakes are high and effective use requires judgment about *when* an algorithmic suggestion is appropriate: in regulated financial advisory, knowing which AI suggestion to deploy is itself a form of human capital, and it correlates strongly with broker seniority.

Third, do brokers learn from using the AI tool and improve their underlying advisory skill? I find evidence consistent with within-broker upskilling that transfers to interactions in which the AI is not active. Comparing brokers in the top and bottom quintiles of treatment share, control customers served by high-exposure brokers convert at 4.36 % on the fifteen-day margin, whereas control customers served by low-exposure brokers convert at only 3.62 %. The 0.74-percentage-point gap is statistically significant ($p \approx 0.024$) despite identical ex ante randomization. AI-suggested phrasing, product framings, and conversational pacing carry over into the broker's interactions with non-treated customers, where the AI tool is not active. The intent-to-treat estimate, which compares treated and control customers *within* the same brokers, therefore understates the full effect of AI deployment, because the control benchmark itself has been lifted by spillover.

Fourth, does using AI more translate into extracting more value from it? The findings indicate that adoption frequency and per-use value diverge sharply across broker ranks. Adoption is measured directly: the platform records four actions a broker can take on each AI suggestion: (i) one-click full adoption, (ii) partial adoption after deleting part of the suggestion, (iii) partial adoption after editing, and (iv) copy as inspiration without sending. Total active adoption (the sum of all four modes) declines monotonically with broker seniority, averaging 27.2 % at L1 and falling to roughly 10 % at L4+, a nearly threefold gradient. Each individual mode contributes to the decline, although the partial-adoption-after-edits mode is not strictly monotone across all five rank levels. Yet the per-use revenue return runs in the *opposite* direction, rising monotonically with seniority (Finding 2). Junior brokers adopt AI suggestions across all four modes at higher rates but extract lower revenue per use; senior brokers adopt selectively and extract substantially higher value when they do. The workers who use AI most are not the workers who gain most from it.

This paper contributes to five strands of literature.

Generative AI and productivity. A growing body of randomized field experiments has converged on a robust, sizeable productivity gain from generative AI in white-collar work. Brynjolfsson et al. (2025) document a 14 % gain in customer support; Cui et al. (2026) pool three Copilot RCTs with nearly five thousand software developers and report a 26 % increase in completed tasks; Noy and Zhang (2023) document 40 % time savings in writing tasks; and Dell'Acqua et al. (2023) report task-conditional gains

for BCG consultants. Adjacent observational and field-deployment evidence (Peng et al., 2023; Hoffmann et al., 2024; Jaffe et al., 2024) corroborates magnitude and persistence, and survey evidence from corporate executives indicates that the gains are concentrated in finance and high-skill services (Baslandze et al., 2026). The settings span customer service, coding, writing, and consulting; in each, the cognitive task is predominantly linguistic or algorithmic, per-decision stakes are modest, and the AI sits inside an internal task rather than directly mediating a customer-facing transaction.

This paper provides, to my knowledge, the first causal evidence on generative AI in a high-stakes professional-services setting deployed in a live commercial environment. The outcome is realized insurance premium revenue from real customer transactions, not a productivity proxy; the AI sits behind a licensed advisor in regulated, multi-thousand-dollar conversations rather than inside an internal workflow; and the broker carries the legal and reputational responsibility for what reaches the customer.

Financial advice and the role of advisors. Financial advice intermediates trillions of dollars of household savings, and a long literature documents that the market is shaped by conflicts of interest, regulatory choices, and heterogeneity in advisor quality. Inderst and Ottaviani (2012) provide a foundational framework for advice provision under commission-based incentives. Empirical work documents systematic distortions in advice. Brokers steer clients toward higher-fee or dominated products (Bergstresser et al., 2009; Christoffersen et al., 2013; Mullainathan et al., 2012; Egan, 2019; Anagol et al., 2017; Guiso et al., 2022); advisor misconduct is widespread and predictable across firms (Egan et al., 2019, 2025); advisors' *own* mistaken beliefs shape the advice they give (Linnainmaa et al., 2021); and women systematically receive different advice from human advisors (Bucher-Koenen et al., 2025; Bhattacharya et al., 2024). Regulatory variation, in particular fiduciary duty, has become a central object of recent quasi-experimental and structural work. Egan et al. (2022) document that fiduciary duty changes broker steering in variable annuities, and Charoenwong et al. (2019) show that the regulatory agency in charge matters for advice quality. Bhattacharya et al. (2025) estimate, in a structural model of entry and advice provision, that fiduciary duty raises risk-adjusted returns by 25 basis points but contracts entry by 16 %. A parallel literature evaluates technological alternatives to human advisors: robo-advisors increase diversification and welfare for less-sophisticated investors (D'Acunto et al., 2019; Rossi and Utkus, 2024; Reher and Sokolinski, 2024), but clients value human advisors for soft skills that automation does not replicate (Greig et al., 2025).

Generative AI is being adopted across financial-advice channels at unusual speed, but the empirical evidence on its effects is thin. Huang et al. (2025) document forecast-accuracy gains from GenAI-assisted sell-side equity-analyst reports, the closest financial-services analogue, though they observe text-detected assistance rather

than realized transactions; Choukhmane et al. (2026) simulate consumer-side LLM advice and identify prompt-driven gaps; Yang et al. (2025) show that placing a human banker between the client and the AI raises follow-through through affective rather than informational channels; Fieberg et al. (2025) and Jiang et al. (2025) extend the robo-advisor template with LLM personalization; and Liu (2024) and Cong et al. (2025) provide the nearest analogues in Chinese insurance and real-estate advisory. However, none of these papers observes both suggestion-level adoption and realized financial outcomes in a regulated advisor-mediated workflow. This paper closes that gap and lets the AI's effect on realized premium revenue be traced directly through each broker's individual adoption choice.

Which workers gain from AI. Whether AI compresses or widens the productivity distribution is a first-order question for the future shape of the labor market. Two mechanisms compete in theory. When AI *substitutes* for tacit skill (encoding as machine-readable instructions what high-ability workers previously held as private know-how), it disproportionately helps low-skill workers and compresses the productivity distribution. When AI *complements* tacit skill (because skilled workers are better at evaluating, editing, and selectively deploying AI output), it disproportionately helps high-skill workers and widens the distribution. The Acemoglu–Restrepo task-displacement framework (Acemoglu and Restrepo, 2018, 2019) embeds both possibilities; Autor (2024) stresses the substitution channel for occupations bottlenecked by access to expert judgment, while Humlum and Vestergaard (2025) note that adoption barriers can amplify pre-existing inequality even when conditional gains are leveling, and de Souza (2026) shows with administrative software-registry data that AI's distributional effects are sector-dependent, hollowing out the office wage distribution while lifting low-skill production wages.

The dominant empirical pattern in field experiments is leveling. Brynjolfsson et al. (2025) document a 14 % average gain in customer support driven mostly by low-skill agents (+34 %); Noy and Zhang (2023) report compression in writing; Cui et al. (2026) find that less-experienced developers gain more in coding; Kanazawa et al. (2026) show AI navigation benefits only low-skilled taxi drivers; and Huang et al. (2025) find the accuracy gain from GenAI-assisted analyst reports is larger for lower-baseline-skill analysts, the leveling pattern holding even within financial services. Two papers are sometimes read as counterexamples, but their amplification claims are narrow. Otis et al. (2024) find $\pm 0.2\sigma$ effects of opposite sign in unregulated Kenyan small-business entrepreneurship, an order of magnitude smaller than the gradient documented here. Dell'Acqua et al. (2023) document a task-conditional "jagged frontier" for BCG consultants in which inside-frontier tasks actually show *compression* across consultant skill, a task-fit result, not a worker-skill amplification result.

This paper documents amplification in a setting that places unusual weight on the human's experience and judgment: regulated, revenue-bearing financial advisory in which the broker must diagnose customer needs, integrate product detail, and exercise judgment about when an AI suggestion is appropriate. The effects are monotone in seniority and complementary rather than substitutive: AI raises productivity most where the broker's residual judgment can do the most work. Where leveling dominates in customer support, writing, coding, and taxi navigation, complementarity dominates here.

AI-induced learning and human capital. Now that productionized generative-AI tools have been in deployment for several years, whether their productivity gains accrue only to contemporaneous AI users or transfer durably as upskilling is a key open question for the long-run impact of the technology, and systematic evidence remains thin. Brynjolfsson et al. (2025) interpret their leveling result as AI encoding the tacit knowledge of top performers and disseminating it to novices. Related work shows that machine-learned tips improve human sequential decisions (Bastani et al., 2025), that human judgment is best elicited as a private-information adjustment (Ibrahim et al., 2021), that proximity to expert coworkers raises junior productivity (Emanuel et al., 2023), and that managerial-tool adoption depends on worker-level absorptive capacity (Bloom et al., 2013). Most existing field RCTs cannot identify whether an individual worker has *learned* from the AI: they observe only the across-worker distribution of gains. Exploiting cross-broker variation in cumulative AI exposure generated by the customer-level randomization, this paper documents the within-broker case directly. Control customers served by high-exposure brokers convert at a significantly higher rate than control customers of low-exposure brokers, despite identical ex ante randomization (Section 4). The counterfactual itself rises with cumulative AI exposure, the field-RCT footprint of the BLR knowledge-dissemination mechanism operating *within* an individual high-skill worker.

Human-AI collaboration and algorithmic advice. AI is being deployed alongside human decision-makers across an expanding range of domains, and a corresponding literature has begun to ask how to measure the effect of adoption and how to design optimal human-AI collaboration mechanisms. Angelova et al. (2025) show that only the top decile of NYC bail judges improves on the algorithm when overriding it, and that the remaining 90 % of overrides are essentially random; Agarwal et al. (2025) develop a sufficient-statistic framework in which humans under-respond to AI predictions because they are overconfident in their own signal rather than under-confident in the AI; and Agarwal et al. (2023) document that radiologists fail to update their beliefs in response to AI diagnostic predictions. Dell'Acqua (2022) shows that confident AI

predictions can crowd out human effort entirely ("falling asleep at the wheel"); the behavioral foundation in algorithm aversion (Dietvorst et al., 2015, 2018) and algorithm appreciation (Logg et al., 2019) predicts that the level of adoption depends on expertise, modifiability, and recent error history, and Skitka et al. (1999) adds the symmetric concern of automation bias. Methodologically, the algorithmic-triage literature (Raghu et al., 2019; Mozannar and Sontag, 2020; McLaughlin and Spiess, 2024) treats human accuracy as fixed and asks which cases to delegate; the meta-analysis of Vaccaro et al. (2024) finds that the average human-AI team underperforms the better of human or AI alone; and Augenblick et al. (2025) provide the QJE-level belief-updating primitive (under-response to strong signals) consistent with the seniority pattern documented here.

Most of this work observes only the aggregate treatment effect or splits gains by ability tier, without separately identifying the *frequency* and the *return* margins of AI use. With suggestion-level adoption logs, this paper decomposes the ITT into adoption frequency $\hat{\pi}$ and per-use LATE and shows that the two move in *opposite* directions across broker ranks: junior brokers adopt more often but extract lower per-use return, while senior brokers adopt selectively and extract substantially higher per-use return. The pattern is the field-setting analogue of Angelova et al. (2025)'s bail-judge result and the empirical footprint of Agarwal et al. (2025)'s overconfidence mechanism: high-ability humans use AI less often but better, exactly because they correctly down-weight AI when their own signal is informative.

Outline. The remainder of the paper is organized as follows. Section 2 describes the institutional setting, the experimental design, and the data. Section 3 reports the average treatment effects on conversion, premium revenue, and engagement, and presents robustness checks (Section 3.4). Section 4 documents the within-broker upskilling result, showing that cumulative AI exposure shifts gains from the extensive to the intensive margin. Section 5 documents customer-segment heterogeneity and the engagement, discovery, and high-value mechanisms. Section 6 documents broker heterogeneity by seniority and the inverse adoption-frequency-versus-return pattern. Section 7 concludes.

2 Institutional Setting, Experimental Design, and Data

This section describes the empirical setting in three parts. I first introduce Ant Insurance and the broker-customer advisory process; then describe the production-grade generative-AI tool and the customer-level randomization that defines the experiment; and finally document the data sources, sample construction, and balance across the treatment and control groups.

2.1 Ant Insurance and China's insurance intermediary market

China's insurance industry collected 5.7 trillion RMB in premiums in 2024. Of that, 962 billion RMB, approximately 17 %, was intermediated by specialized agencies and brokerages, across 2,162 licensed firms. Ant Insurance is the dominant online insurance brokerage in this market and has ranked first among all intermediaries since 2022. In 2024 it generated 9.43 billion RMB in revenue, representing approximately 5.9 % of the total intermediary market and roughly equal to the second and third largest firms combined; its estimated premium volume stands at 55–70 billion RMB, or 6–7 % of the channel. The platform connects over 600 million customers who have purchased insurance through the Alipay app, with 1,264 products from more than 90 partner insurers spanning health, life, accident, annuity, property, auto, travel, and pet lines.[1]

This paper focuses primarily on the life-insurance and critical-illness lines, the platform's highest-complexity and highest-average-premium products. 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. Average transaction premiums in these lines run into the tens of thousands of RMB, and the most complex policies exceed 200,000 RMB.

2.2 The broker-customer advisory process

Each customer who seeks insurance advice on the platform is matched to a licensed insurance broker by the platform's allocation algorithm. The broker is the agent of record: all product recommendations are sent by the broker, and the customer never interacts with the AI tool directly. The advisory conversation proceeds primarily through the platform's in-house instant-messaging (IM) interface. A typical conversation spans multiple days, with the customer raising questions about coverage terms, health disclosure requirements, premiums, and beneficiaries before making a purchase decision. Conversations that result in a sale average about eighteen dialogue rounds from first contact to purchase; the most complex transactions involve fifty or more exchanges. Figure A1 in the appendix shows the customer-facing screens by which a customer locates a broker, browses broker profiles, initiates the chat, and views recommended products.

Figure 1 illustrates the AI tool's position in the broker's workspace. When a customer sends an IM message, the tool processes it and returns a set of suggested replies and candidate products to the broker's panel. The broker then decides whether to use the suggestion, modify it, or write an independent reply. This interaction is invisible

[1]Source: *Yearbook of China's Insurance* (2025). Profit figure: approximately 694 million RMB, roughly 15 % of industry profit.

to the customer. When the conversation warrants, the broker may also follow up by telephone; telephone conversations are not mediated by the AI tool.

Each broker carries a five-step seniority rank (L1 through L5) under the platform's internal career system, determined by cumulative performance, training completion, and tenure. The rank is updated periodically and serves as a coarse but public signal of broker quality. For statistical power, I collapse the two senior ranks into a single group (L4+), yielding four seniority levels: L1 (most junior), L2, L3, and L4+ (most senior). L2 dominates the analysis sample (55.9 % of brokers), with L1, L3, and L4+ accounting for 12.5 %, 20.4 %, and 11.2 % respectively. Median broker tenure is approximately ten months (p75 eighteen months; longest-tenured broker, 3.7 years), reflecting that the insurance brokerage platform was launched roughly three years before the experiment and is in a phase of rapid growth. Top-decile brokers produce roughly eight to ten times the realized premium revenue of bottom-decile brokers, so the dispersion in broker ability is economically large.

2.3 The generative-AI tool

The platform's AI tool, *Tongtiandai*, combines three components: a fine-tuned large language model (Qwen2.5-72B, comparable in scale to GPT-4o or Claude 3.5 Sonnet), a curated insurance knowledge base covering product terms and regulatory requirements, and a corpus of past advisory transcripts. For each incoming customer message, the tool conditions on the message text together with the customer's personal information, age, gender, city, family composition, and past insurance activity, and returns a set of suggested replies and candidate products to the broker (Figure 1). Figure A2 in the appendix shows screen captures of the broker workstation with the AI tool active during a live conversation.

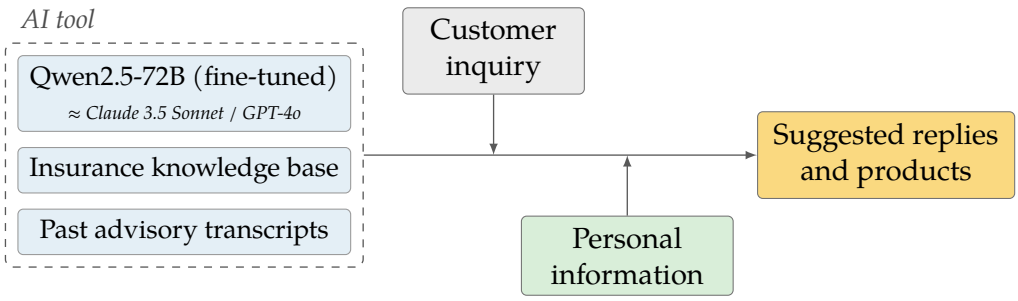


Figure 1: How the AI tool generates suggestions for the broker

Notes: For each incoming customer message, the 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 inquiry and personal information (age, gender, city, family composition, past insurance activity) and returns suggested replies and candidate products to the broker. The customer never sees the AI tool; all messages reaching the customer are sent by the broker.

When the broker receives a suggestion, four handling actions are recorded: fully

adopt and send, send after partial deletion, send after edits, or copy as inspiration without sending in this conversation. In aggregate across the treatment sample, roughly 3.6 % of suggestions are fully adopted, 8.5 % are partially adopted with a usable subset retained, 0.3 % are sent after edits, and 3.3 % are copied as inspiration. These adoption rates vary substantially across broker seniority levels, as documented in Section 5.

2.4 Randomization and design

The experiment ran for twenty calendar days in mid-2025 and randomized at the customer level: every new customer matched to a broker was independently assigned to the treatment group, in which the AI tool was activated in the broker's workspace for that conversation, or to the control group, in which the AI was disabled. The same broker typically handled both groups during the same week. There is no broker-level randomization and no customer-level self-selection into treatment; assignment is determined by the platform's Darwin A/B system before the broker is assigned.

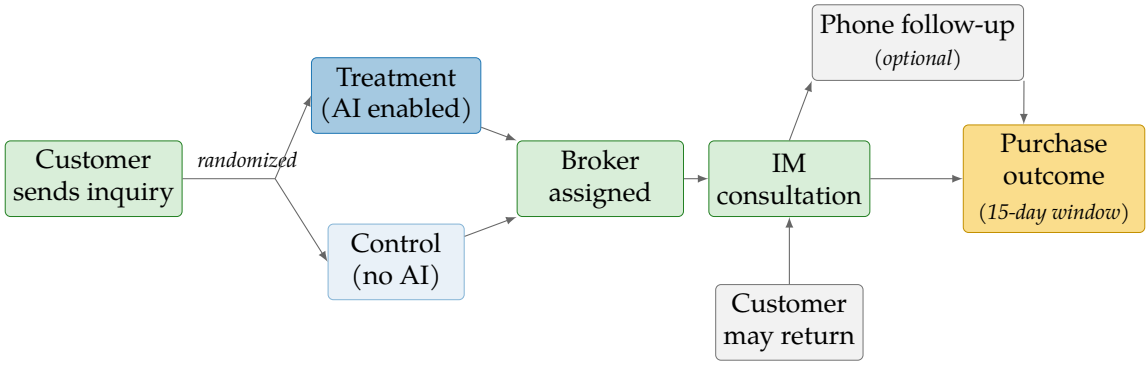


Figure 2: Customer-level randomization and outcome window

Notes: Each new customer arriving at the platform is independently randomized to the treatment group (AI enabled) or control group (no AI) before a broker is assigned. Both the treatment assignment and the assigned broker are held fixed for a given customer throughout the experimental window. The customer's purchase outcome is tracked for fifteen days after first contact.

A treated assignment *enables* AI in the broker's workspace for that conversation but does not *force* it: the broker may consult the suggestions, use them throughout, or ignore them entirely. I treat actual AI use as an endogenous behavioral choice and instrument it with random assignment in the IV specifications of Section 3. The brokers did not know which customers were assigned to the treatment or control group; the platform interface displays the same workspace layout regardless of assignment, with AI suggestions appearing only when the tool is activated. This rules out differential broker effort toward treatment-group customers.

Identification assumptions. The experiment's causal identification rests on three conditions. (1) *Random assignment:* treatment is determined independently of cus-

tomer characteristics, broker characteristics, and the unobserved determinants of insurance purchase, verified by the balance table in Section 2.5. (2) *Exclusion:* AI assignment affects the purchase outcome only through the broker's use of AI suggestions in the IM conversation; in particular, it does not affect the broker's behavior in telephone follow-up calls, where AI is not present. (3) *SUTVA* (*Stable Unit Treatment Value Assumption*)*:* each customer's potential outcomes depend only on her own treatment assignment, not on the assignments of other customers. The standard threat to SUTVA in this setting is broker-mediated spillover: a broker who serves both treatment and control customers in the same week may learn from AI-assisted conversations and apply that learning to control customers, attenuating the ITT estimate. The ITT estimate is therefore a lower bound on the effect that would obtain from a full rollout where every customer is treated.

Outcome window. The fifteen-day outcome window was chosen because it covers essentially all completed transactions in the life and critical-illness lines: the vast majority of customers who purchase do so within a few days of their first conversation, and extending the window beyond fifteen days adds very few additional conversions. There is no right-censoring concern: the outcome is computed ex post as the customer's total premium over the fifteen calendar days following first contact, regardless of when the experimental window closed.

Sample funnel and analysis sample. Figure 3 illustrates the customer funnel from initial broker match through to a fifteen-day purchase. Throughout the paper I restrict the analysis sample to customers who entered at least one substantive conversation, defined as having a positive count of dialogue rounds on at least one day in the window.

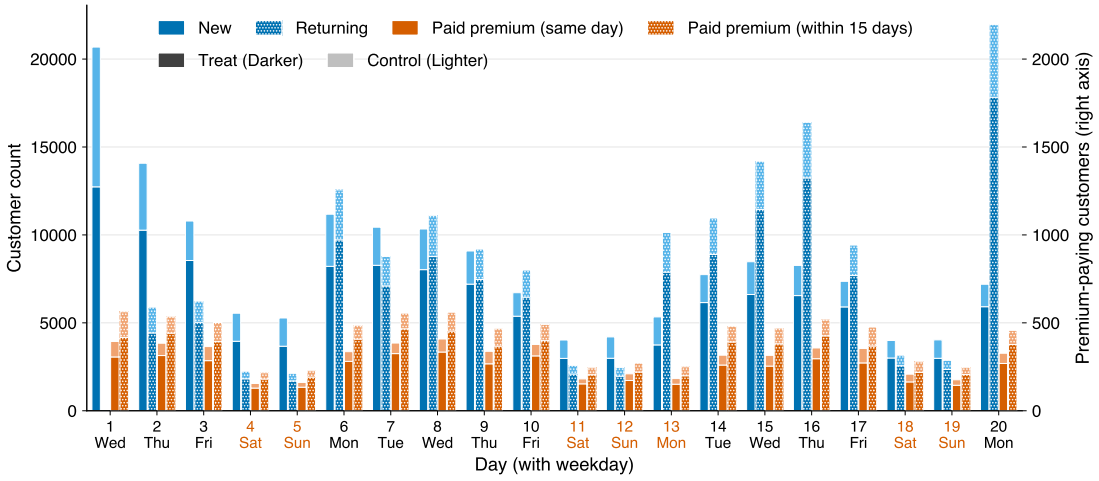


Figure 3: Customer funnel from initial broker match to a 15-day purchase
Notes: Each layer is split between the treatment group (darker) and the control group (lighter). The analysis sample throughout the paper is the "engaged at least once" layer.

The final analysis sample contains 164,658 engaged customers served by 1,063 brokers. Table 1 summarizes the composition.

	Count	Share
Customers (engaged in ≥ 1 conversation)	164,658	
Assigned to treatment (AI enabled)	128,957	78.32 %
Assigned to control	35,701	21.68 %
Brokers serving these customers	1,063	
Customers with same-day purchase	5,280	3.21 %
Customers with 15-day purchase	7,290	4.43 %
Share of treated customers with actual AI use ($\hat{\pi}$)		44.3 %

Table 1: Analysis sample

Notes: Share for treated/control rows is relative to total engaged customers. Share for purchase rows is the 15-day conversion rate among all engaged customers.

Customer segments and broker ranks. Each customer carries a four-step pre-experiment value segment inferred from prior platform interactions, ordered B (lowest expected lifetime value) < A < S < S+ (highest); customers without historical interactions are reported as a separate "New" group. Table 2 reports the joint distribution of customers across broker rank and customer segment.

Broker rank	Customer segment					Customers (share)	Brokers (share)
	B	A	S	S+	New		
L1	5,760	6,298	562	702	13	13,335 (8.1 %)	132 (12.5 %)
	43.2 %	*47.2 %*	*4.2 %*	*5.3 %*	*0.1 %*		
L2	35,068	47,112	4,240	4,414	848	91,682 (55.7 %)	594 (55.9 %)
	38.3 %	*51.4 %*	*4.6 %*	*4.8 %*	*0.9 %*		
L3	13,770	18,362	1,962	1,997	1,707	37,798 (23.0 %)	217 (20.4 %)
	36.4 %	*48.6 %*	*5.2 %*	*5.3 %*	*4.5 %*		
L4+	6,973	10,099	1,314	1,614	1,843	21,843 (13.3 %)	119 (11.2 %)
	31.9 %	*46.2 %*	*6.0 %*	*7.4 %*	*8.4 %*		
Full sample	61,571	81,871	8,078	8,727	4,411	164,658	1,063
	37.4 %	*49.7 %*	*4.9 %*	*5.3 %*	*2.7 %*		

Table 2: Customer-segment composition by broker rank

Notes: Top number in each cell is the customer count in the broker rank × customer segment; the italic number below is the within-rank share across customer segments (percentages sum to 100 % within each row). The two rightmost columns report the customer count (and percentage of total customers) and the broker count (and percentage of total brokers) at each rank. B and A together account for 87.1 % of customers; S and S+ together cover 10.2 %. The share of S+ and New customers rises markedly with broker seniority, reflecting the platform's practice of routing harder cases to more experienced brokers.

Panel A of Table 2 shows that customers concentrate in the lower-value segments: B and A together account for 87.1 % of the sample, while S and S+ together cover 10.2 %.

Panel B reveals that the assignment of high-value and new customers to brokers is non-random: the share of S+ and New customers rises markedly with broker seniority, reflecting the platform's practice of routing harder cases to more experienced brokers. The L4+ group allocates 7.4 % of its customers to S+ and 8.4 % to New, compared to 5.3 % and 0.1 % at L1.

2.5 Data description

The analysis draws on three linked datasets from the Ant Open Research Laboratory sandbox, covering all customer-broker interactions during the twenty-day experimental window.

Customer-broker-day dataset. The primary dataset records, for each customer-broker pairing, the daily count of dialogue rounds, same-day and fifteen-day realized premium, the number of AI suggestions generated and adopted (in the treatment group), and customer characteristics including age and a pre-experiment value segment.

Broker-day dataset. A daily panel of broker-level outcomes covers service time (IM and telephone duration), product-level conversion counts and premium, big-ticket policy counts, and daily AI adoption totals broken down by action type (fully adopted, partially adopted, or copied as inspiration).

In addition, broker-level static information is available for each broker in the analysis sample, including seniority rank (L1–L5), joining date (used to compute tenure), and an ability-score rating.

Table 3 reports descriptive statistics for the analysis sample. Panel A describes premium outcomes conditional on a positive purchase: the fifteen-day premium distribution is highly right-skewed (median around 2,000 RMB, P90 around 60,000 RMB), motivating the use of Poisson pseudo-maximum-likelihood (PPML) for combined-effect estimates. To limit the influence of extreme outliers, I winsorize the premium variables at the 99th percentile of the positive subsample and use the winsorized series as the baseline outcome throughout the analysis; the unwinsorized series serves as a robustness check. Panel B describes dialogue-round outcomes. Panel C reports the unconditional purchase probability across all customers (4.4 %). Panel D summarizes AI exposure in the treated subsample: 47.2 % of treated customers had at least one AI suggestion generated for their conversation, with a peak daily count of 2.72 suggestions on average.

Variable	N	Mean	SD	P10	P25	P50	P75	P90
Panel A: Premium outcomes, conditional on positive purchase (RMB)								
15-day premium	7,290	27,078	96,800	100	392	1,998	17,154	59,874
winsorized at p99	7,290	23,332	56,000					
Same-day premium	5,280	36,300	127,500	125	497	2,981	24,116	99,648
winsorized at p99	5,280	31,734	60,500					
Panel B: Dialogue-round outcomes								
Cumulative dialogue rounds (20-day)	164,658	5.35	10.76	1	1	3	5	11
Rounds before first purchase (purchasers)	5,280	18.0	24.1	1	4	11	23	41
Panel C: Purchase indicators (full sample)								
Customer purchases within 15 days	164,658	0.0443	0.2057	0	0	0	0	0
Customer purchases on the same day	164,658	0.0321	0.1762	0	0	0	0	0
Panel D: AI exposure (treated subsample)								
Any AI suggestion shown for customer	129,006	0.472	0.499	0	0	0	1	1
Peak daily AI suggestions for customer	129,006	2.72	6.75	0	0	0	2	7

Table 3: Summary statistics for the analysis sample

Notes: Premium variables in Panel A are reported on the purchaser subsample only; percentiles are exact. Winsorization at p99 mechanically leaves all reported percentiles ($\leq$ P90) unchanged, so they are omitted from those rows. Panel C reports the unconditional Bernoulli rate of purchase on the full sample. Panel D reports AI exposure variables on the treated subsample (129,006 customers); control-group conversations do not invoke the AI tool by design.

2.6 Balance check

Random assignment at the customer level produces well-balanced treatment and control groups. Treatment shares lie between 78.0 % and 79.2 % within every broker rank and every customer segment (Table 4). The one exception is the "New" segment, which is 76.3 % treated, consistent with sampling noise given its small size. Unconditional two-sample t-tests on the headline outcomes confirm that treatment-control differences are statistically significant even before any covariate adjustment.

Broker rank	Customer segment					Row marginal
	B	A	S	S+	New	
L1	78.92 %	79.58 %	78.51 %	77.60 %	53.85 %	79.12 %
L2	78.04 %	78.40 %	78.28 %	78.07 %	75.71 %	78.22 %
L3	78.39 %	78.22 %	77.23 %	79.24 %	74.34 %	78.11 %
L4+	77.11 %	79.17 %	80.56 %	76.79 %	77.63 %	78.34 %
Col. marginal	78.14 %	78.53 %	78.43 %	78.04 %	76.26 %	78.29 %

Table 4: Treatment share by broker rank × customer segment

Notes: Each cell reports the share (%) of customers in that broker rank × customer segment cell who were assigned to the treatment group. The bottom-right cell (78.29 %) is the full-sample treatment share. Marginal shares lie between 78.0 % and 79.2 % in every broker rank and every segment except "New" (76.26 %), consistent with random assignment. The L1×New cell ($N = 13$; 53.85 %) reflects sampling noise in a very small cell.

This study was remotely conducted on the *Ant Open Research Laboratory*[2] 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 authors can only remotely conduct empirical analysis and individual observations are invisible.

3 Empirical Strategy and Main Results

This section first sets out the estimating equations and the identification strategy that exploits customer-level randomization, then reports the average treatment effects of AI assignment on premium revenue, conversion, and customer engagement, and finally verifies that the results are robust to specification, estimator, clustering, and sample choices.

3.1 Empirical Strategy

The customer-level analysis tracks three outcomes:

- **Premium-15d**: realized premium received within fifteen days of first contact, winsorized at the 99th percentile of the positive subsample. This is the headline measure of realized revenue.
- **Premium-1d**: cumulative same-day premium, winsorized. A robustness measure capturing faster conversions.
- **Dialogue rounds**: cumulative count of broker-customer message exchanges across all twenty days. By sample construction this is bounded below by one, so the extensive margin is mechanically equal to one and is not estimated separately. A second engagement outcome, pre-buy rounds, is introduced when needed in Section 3.3.

Table 5 reports the unconditional treatment-control differences on these outcomes. AI assignment raises every measured outcome, with all conversion-rate and dialogue-round differences significant at the 1 % level.

[2] https://www.deor.org.cn/labstore/laboratory

Outcome	Control	Treated	Diff	Pct. diff	p-value
15-day outcomes					
Conversion rate Pr(prem15 > 0)	4.11 %	4.51 %	0.40 pp	+9.7 %	<0.001
15-day premium (RMB)	1,009.70	1,251.67	241.97	+24.0 %	0.024
15-day premium, winsorized at p99	907.46	1,068.33	160.87	+17.7 %	0.031
Same-day outcomes					
Conversion rate Pr(prem1 > 0)	2.90 %	3.29 %	0.39 pp	+13.4 %	<0.001
Same-day premium (RMB)	967.41	1,220.02	252.61	+26.1 %	0.027
Same-day premium, winsorized at p99	905.51	1,050.06	144.55	+16.0 %	0.093
Dialogue-round outcomes					
Dialogue rounds (20-day cumulative)	5.08	5.42	0.34	+6.7 %	<0.001
Pre-buy rounds (full sample)	0.523	0.592	0.069	+13.2 %	0.025

Table 5: Unconditional treatment-control differences on outcome variables
Notes: p-values from Welch two-sample t-tests. Diff = Treated − Control; Pct. diff = (Treated − Control) / Control. Conversion-rate diffs are reported in percentage points (pp); premium and dialogue-round diffs are in their native units.

The treatment indicator Treat_i equals one if customer i is randomly assigned to the AI-enabled group. The endogenous indicator of actual AI use, UseAI_i (`expo_d`), equals one if any AI suggestion was shown to the broker during the customer's conversation. The control vector X_i comprises customer-segment fixed effects, first-day (calendar date of first contact) fixed effects, broker-rank fixed effects, and broker-hire-cohort (calendar-quarter) fixed effects. Standard errors are clustered at the broker level (∼1,063 clusters) throughout. For each outcome y_i I estimate up to five specifications.

PPML. The headline estimator for premium revenue is Poisson pseudo-maximum-likelihood on the level of the outcome:

$$E[y_i \mid \text{Treat}_i, X_i] \;=\; \exp(\alpha + \beta\,\text{Treat}_i + X_i'\gamma). \tag{1}$$

The coefficient β yields a percent effect $100(e^{\beta} - 1)$. PPML is the natural estimator for premium revenue: it accommodates the mass of zeros (most customers do not purchase) and the right-skewness of the positive subsample, and it consistently estimates the conditional mean even when the outcome is heteroscedastic in a multiplicative way.

Extensive margin. A linear probability model on the conversion indicator $\mathbf{1}\{y_i > 0\}$:

$$\mathbf{1}\{y_i > 0\} \;=\; \alpha + \beta_E\,\text{Treat}_i + X_i'\gamma + \varepsilon_i, \tag{2}$$

with β_E expressed in percentage points. I report two variants of equation (2). The pooled LPM identifies the ITT from the customer-level randomization without adjusting for broker identity. The broker-FE LPM replaces the constant α with a broker fixed

effect $\alpha_{b(i)}$ (where $b(i)$ is the broker assigned to customer i) and identifies the same ITT from comparing treated and control customers served by the *same* broker, absorbing all broker-level heterogeneity (skill, effort, hours, customer mix) as a fixed effect. Under customer-level randomization the two variants target the same population parameter; comparing them tests whether cross-broker variation in treat share happens to correlate with cross-broker variation in conversion. A small gap (as observed) indicates that the pooled ITT is not driven by such cross-broker composition effects.

IV-Extensive. The treatment indicator captures *assignment* to AI, not actual use. Compliance is partial: brokers deploy AI suggestions in roughly 44 % of treated conversations. To recover the effect of actual use, I estimate two-stage least squares on the conversion indicator, instrumenting UseAI_i with Treat_i. The first stage is

$$\text{UseAI}_i \;=\; \pi_0 + \pi_1 \,\text{Treat}_i + X_i'\delta + u_i, \tag{3}$$

and the second stage is

$$\mathbf{1}\{y_i > 0\} \;=\; \alpha + \beta_{\text{LATE}}\,\widehat{\text{UseAI}}_i + X_i'\gamma + v_i. \tag{4}$$

The coefficient β_{LATE} is the local average treatment effect on conversations in which the broker actually deployed AI. The exclusion restriction is mechanical: in control conversations $\text{Treat}_i = 0$ disables the AI tool entirely, so Treat_i can affect the conversion indicator only through UseAI_i. The IV rescales the ITT by $1/\hat{\pi}_1 \approx 1/0.443 \approx 2.3$.

Intensive margin. On the purchaser subsample ($y_i > 0$),

$$\log y_i \;=\; \alpha + \beta_I \,\text{Treat}_i + X_i'\gamma + \varepsilon_i, \tag{5}$$

captures changes in the per-buyer outcome conditional on conversion. This is the intensive-margin counterpart to equation (2).

3.2 Premium outcomes

Table 6 reports the pooled ITT estimates on premium outcomes. The PPML estimate on the fifteen-day premium is +30.7 % ($p < 0.01$), with a nearly identical estimate of +30.2 % for the same-day premium. These are large gains by any standard: the causal productivity literature reports effects of 14 % for customer support (Brynjolfsson et al., 2025) and 14 % for writing tasks (Noy and Zhang, 2023), and the present setting involves an order of magnitude higher transaction value per customer.

Outcome	Model	β	SE(β)	Effect	SE(eff)	
Premium-15d	PPML	0.268	0.078	+30.72 %	10.20	***
	Extensive (LPM)	0.006	0.001	+0.63 pp	0.12	***
	Extensive Broker FE	0.004	0.001	+0.43 pp	0.12	***
	IV-Extensive (2SLS)	0.014	0.003	+1.41 pp	0.28	***
	Int log (OLS)	0.115	0.069	+12.21 %	7.75	*
Premium-1d	PPML	0.264	0.092	+30.23 %	12.01	***
	Extensive (LPM)	0.006	0.001	+0.58 pp	0.10	***
	Extensive Broker FE	0.004	0.001	+0.41 pp	0.10	***
	IV-Extensive (2SLS)	0.013	0.002	+1.32 pp	0.24	***
	Int log (OLS)	0.154	0.082	+16.68 %	9.63	*

Table 6: Pooled treatment effects on premium outcomes

Notes: N =164,658 customers and 1,063 brokers. PPML and Int log effects expressed as percent changes $100(e^{\beta}-1)$; Extensive, Extensive Broker FE, and IV-Extensive effects in percentage points (pp). Extensive Broker FE is a within-broker LPM. IV-Extensive instruments actual AI use (`expo_d`) with random assignment (`Treat`); first-stage F = 14,103; compliance $\hat{\pi}$ = 44.3 %. Int log is OLS on log(Premium) restricted to the purchaser subsample (N = 7,290 for 15d, N = 5,280 for 1d). Standard errors clustered at the broker level. Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Controls: customer-segment FE, first-day FE, broker-rank FE, broker-cohort FE.

Extensive margin. The extensive-margin (conversion rate) lift is +0.63 pp ($p < 0.01$), a 16 % relative increase on a control-group conversion rate of 3.8 %. The within-broker estimate, identified from comparing customers served by the *same* broker, is +0.43 pp, indicating that roughly two-thirds of the extensive-margin lift survives within-broker variation, consistent with the randomization producing balance within as well as across brokers.

Use vs. assignment. The compliance rate is 44.3%: in fewer than half of treated conversations did the broker actually deploy the AI tool. The IV-Extensive estimate of +1.41 pp is the local average treatment effect (LATE) on conversations in which the broker chose to use AI. This exceeds the ITT by a factor of $1.41/0.63 \approx 2.3\times$, consistent with the identity ITT $= \hat{\pi} \times$ LATE ($0.443 \times 1.41 \approx 0.63$ pp): the effect of AI assignment operates *through* actual AI use, not merely through exposure. Treated conversations in which the broker did not use AI are functionally indistinguishable from the control group. The first-stage F statistic of 14,103 far exceeds conventional weak-instrument thresholds, so the IV is strong.

Intensive margin. Conditional on purchase, treated customers pay approximately 12 % more (Int log: +12.2 % on Premium-15d, significant at the 10 % level). The relatively wide standard errors on the intensive-margin estimates reflect the small size of the purchaser subsample (7,290 out of 164,658), but the positive direction is consistent across both the fifteen-day and same-day windows.

3.3 Customer engagement

To trace the mechanism behind the conversion gain, I look at two engagement outcomes. The first, **dialogue rounds**, is the cumulative count over the twenty-day window (introduced in Section 3.1). The second, **pre-buy rounds**, is the number of dialogue rounds from first contact through the day of the first transaction, defined only on the purchaser subsample of 5,280 customers; it measures the length of the closing path conditional on a purchase occurring. Together, the two outcomes distinguish "AI sustains engagement and brings additional customers across the line" from "AI accelerates the closing path for customers who were already converting."

Table 7 reports effects on dialogue engagement.

Outcome	Model	β	SE(β)	Effect	SE(eff)	
Dialogue rounds	PPML	0.086	0.013	+9.02 %	1.39	***
	Int log	0.059	0.006	+6.07 %	0.66	***
Pre-buy rounds	PPML	−0.011	0.046	−1.09 %	4.59	

Table 7: Pooled treatment effects on dialogue rounds

Notes: Dialogue rounds is cumulative dialogue rounds over the twenty-day window (full sample, $N = 164{,}658$). Pre-buy rounds is cumulative dialogue rounds from first contact through the day of first purchase, defined on the purchaser subsample ($N = 5{,}280$). Effects in percent changes $100(e^{\beta} - 1)$. Standard errors clustered at the broker level. Controls as in Table 6.

Dialogue rounds. AI assignment increases cumulative dialogue rounds over the twenty-day window by approximately 9 % (PPML) or 6 % (log-OLS). Treated conversations are *longer*, not shorter. This finding immediately rules out one popular narrative about AI: that it saves time by accelerating the conversation to a close. In this setting, AI sustains and expands customer engagement rather than compressing it.

Pre-buy rounds. The number of dialogue rounds from first contact to purchase, conditional on purchase, is essentially unchanged: the PPML estimate is −1.1 % (not significant). AI does not shorten the closing path for customers who were already on a converting trajectory. Instead, the conversion-rate gain reflects AI bringing *additional customers*, those who would not have converted without AI, far enough through the conversation to make a purchase decision.

Together, the dialogue-round findings suggest a specific mechanism: AI sustains customer engagement in conversations that would otherwise stall, enabling brokers to deliver the information customers need before they disengage. The average dialogue-rounds gain (+9%) and the zero pre-buy-rounds change are quantitatively consistent with this interpretation, and are supported by the heterogeneity evidence in Section 5.2.

3.4 Robustness

3.4.1 Specification stability

Table 8 reports the headline ITT estimates under progressive addition of controls. I start from the raw treatment-control difference (no controls), then sequentially add broker-rank fixed effects, customer first-day fixed effects, and broker-hire-cohort fixed effects. Because treatment is randomized at the customer level, adding controls should not substantially change the point estimates, and it does not. The fifteen-day LPM estimate moves from +1.80 pp (baseline) to +1.69 pp (with broker fixed effects) to +1.80 pp (full controls). The same-day estimates exhibit the same stability.

	(1) Baseline	(2) +Broker FE	(3) +Customer FE	(4) Full controls
Panel A: Same-day conversion, LPM (pp)				
Treat	1.532***	1.462***	1.615***	1.532***
	(0.037)	(0.038)	(0.037)	(0.037)
N		563,000		
Panel B: 15-day conversion, LPM (pp)				
Treat	1.800***	1.690***	1.800***	1.800***
	(0.041)	(0.042)	(0.041)	(0.041)
N		564,000		
Panel C: Same-day conversion, Logit AME (pp)				
Treat	1.716***	1.711***	1.638***	1.716***
	(0.040)	(0.041)	(0.040)	(0.040)
N		563,000		
Panel D: 15-day premium, PPML (%)				
Treat	+463.8%***	+429.0%***	+379.0%***	+463.8%***
	(46.5)	(57.9)	(68.0)	(46.5)
N		563,669		
Broker FE		✓		✓
Customer FE			✓	✓
Day FE	✓	✓	✓	✓

Table 8: Robustness: progressive control addition

Notes: Point estimates and standard errors (in parentheses) for the treatment indicator under progressive addition of fixed effects. Column (1) is the baseline specification with day fixed effects only; column (2) adds broker fixed effects; column (3) adds customer first-day fixed effects; column (4) includes the full set of controls (broker, customer-day, and broker-hire-cohort fixed effects). Sample sizes are reported once per panel because they are common across the four columns within a panel. Panels A–C are at the customer × day level (so the conversion-rate point estimates are not directly comparable to the customer-level +0.63 pp headline in Table 6; the takeaway here is the stability of the coefficient across columns, not its level). Panel D uses PPML on the *unwinsorized* 15-day premium level; the very large percentage effect reflects right-skew that the winsorized PPML headline of +30.7 % in Table 6 removes. Standard errors clustered at the broker level. Significance: $^{***}p < 0.01$.

3.4.2 Alternative estimators

The extensive-margin effect is robust to the choice of estimator. The linear probability model estimate of +1.53 pp (one-day) and the Logit average marginal effect of +1.72 pp are in close agreement. The PPML and log-OLS specifications for the combined (level) and intensive outcomes produce directionally consistent estimates. Probit replaces Logit with marginal effects within 5 % of the LPM.

3.4.3 Standard error clustering

The headline results cluster standard errors at the broker level (∼1,063 clusters). Clustering at the customer level yields substantially smaller standard errors (customers are less correlated within broker than within customer), so the broker-clustered SE is conservative. Two-way clustering (broker and customer) yields standard errors similar to broker-level clustering. The main results are significant at the 1 % level under all three clustering schemes.

3.4.4 Sample robustness

The raw customer-day table contains five days with ETL-anomalous row counts (approximately 20–200 times the normal daily count, with near-zero compressed file size, consistent with null-heavy snapshot rows). These five days contribute primarily to the full sample count; the engaged-customer analysis sample used throughout the paper is largely unaffected because most of those anomalous rows correspond to zero-engagement entries that are dropped by the engagement filter. Including the anomalous days in the analysis sample (with appropriate flags) does not materially change the main results.

Trimming the top and bottom 1 % of brokers by total customer volume produces ITT estimates within the confidence intervals of the main specification.

4 Broker Learning from AI Exposure

This section asks whether brokers learn from cumulative AI exposure. Section 4.1 compares brokers in the high- and low-exposure tails of the treat-share distribution and shows that AI experience shifts gains from the extensive to the intensive margin and lifts performance even in non-AI interactions, indicating broker upskilling rather than mere reliance on the AI tool. Section 4.2 restricts to the middle 50 % of the treat-share distribution as a robustness check on the pooled main results.

The customer-level randomization fixes each customer's treatment status but, due to finite-sample variation, generates cross-broker variation in the *share* of treated cus-

tomers a broker ultimately serves. This share is exogenous conditional on broker identity, and serves as a natural proxy for cumulative AI exposure during the experiment: a broker who happened to be matched with more treated customers had more occasions to see, evaluate, and either deploy or override AI suggestions.

4.1 Comparing high- and low-exposure brokers

I define two non-overlapping bands at the tails of the treat-share distribution, which is tightly concentrated around the 78 % randomization target (Table A1 in the appendix). The **low-exposure** band (LO) collects brokers in the p5–p25 range of treat share, $[0.703, 0.757]$, and the **high-exposure** band (HI) collects brokers in the p75–p95 range, $[0.812, 0.856]$. The two bands together cover about 40 % of brokers and 40 % of customers in the analysis sample. Figure 4 plots the full cross-broker distribution together with the two bands.

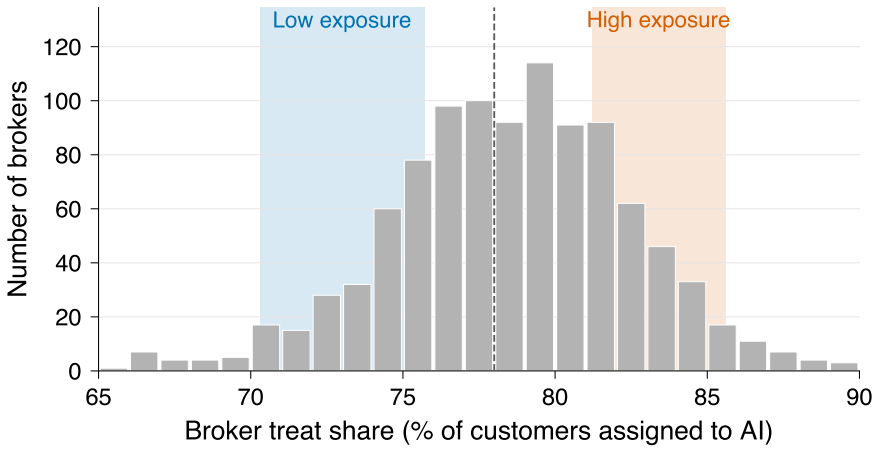


Figure 4: Distribution of broker treat share, with the low- and high-exposure bands
Notes: One bar per one-percentage-point bin across 1,071 brokers. Dashed line: the 78 % randomization target. Shading marks the low-exposure (p5–p25) and high-exposure (p75–p95) bands; axis trimmed to $[65, 90]$.

The two bands are well balanced on observable broker and customer characteristics. Broker-rank composition is similar in both: each band is dominated by L2 brokers (55–60 %), with comparable shares of junior (L1) and senior (L4–L5) brokers, and the customer-segment mix (B, A, S, S+, and New) is also comparable across the two bands. Within each band the treatment share is essentially flat across customer segments, ruling out tier-driven selection as an explanation for the band differences below.

Outcome	Model	Low-exposure (LO)		High-exposure (HI)	
		Effect	SE	Effect	SE
		(1)	(2)	(3)	(4)
Premium-15d	PPML	+46.5%**	25.6	+64.6%***	27.1
	Ext	+1.13 pp***	0.25	+0.14 pp	0.29
	Ext-FE	+0.88 pp***	0.24	+0.02 pp	0.29
	IV-Ext	+2.67 pp***	0.60	+0.31 pp	0.65
	Int log	−2.85%	14.4	+41.7%**	22.9
Premium-1d	PPML	+54.1%**	33.8	+73.4%***	36.2
	Ext	+0.94 pp***	0.22	+0.30 pp	0.24
	Ext-FE	+0.69 pp***	0.21	+0.19 pp	0.24
	IV-Ext	+2.22 pp***	0.51	+0.65 pp	0.55
	Int log	+23.6%	21.8	+56.5%**	29.7
Dialogue rounds	PPML	+11.2%***	2.30	+5.78%***	2.04
	Int log	+8.68%***	1.42	+4.22%***	1.33

Table 9: Treatment effects by broker AI-exposure band

Notes: LO = brokers whose treat share falls below the 25th percentile of the cross-broker distribution (≤ 0.757; 218 brokers, 34,264 customers). HI = brokers above the 75th percentile (≥ 0.812; 201 brokers, 32,741 customers). Effect units: % (PPML, Int log) or pp (Ext, Ext-FE, IV-Ext). Standard errors clustered at the broker level. Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Controls: customer-segment FE, first-day FE, broker-rank FE, broker-cohort FE. The two bands are comparable in broker-rank and customer-tier composition.

Asymmetric effects. Table 9 reveals a striking asymmetry. LO brokers show a large, precisely estimated *extensive-margin* gain: the IV-Extensive LATE on Premium-15d is +2.67 pp ($p < 0.01$), and dialogue volume rises by +11.2 % (PPML, $p < 0.01$). Their intensive margin is near zero and not significant. HI brokers show the opposite pattern: the extensive margin is statistically zero across all specifications, while the intensive margin is large and significant (+41.7 % on Premium-15d, $p < 0.05$; +56.5 % on Premium-1d, $p < 0.05$). PPML on the level of premium is higher for HI than LO (+64.6 % vs. +46.5 %), so the intensive-margin gain dominates in dollar terms for high-exposure brokers.

The control-group conversion gap: spillover or learning. A more telling comparison emerges when I look at the conversion rate of the *control* customers within each band. Because random assignment is at the customer level, control customers served by HI brokers and control customers served by LO brokers should have identical expected conversion rates ex ante. Yet recovered control-group rates differ sharply:

	15-day conversion (p_C)	Same-day conversion (p_C)
LO band	3.62 %	2.58 %
HI band	4.36 %	2.95 %
Δ (HI−LO)	+0.74 pp	+0.37 pp

A two-proportion z-test on the 15-day control-group gap rejects equality at the 5 % level ($z = 2.25$, $p \approx 0.024$). HI brokers' control customers convert about 20 % more often (relative) than LO brokers' control customers, despite being drawn from the same randomization pool. Two non-exclusive channels can produce this gap, both of which are consistent with brokers learning from cumulative AI exposure:

- *Spillover* (*within-broker*): An HI broker, having served many treated customers in the same window, internalizes AI-suggested phrasing, product framings, or conversational pacing and carries them into interactions with control customers as well. The treatment effectively contaminates the control group within the broker.
- *Learning* (*within-broker*): HI brokers accumulate more practice with AI suggestions, raise their general advisory skill, and apply that improved skill uniformly across their caseload, including to control customers.

Implication for interpretation. Both channels imply that the *measured* extensive-margin ITT in the HI band (+0.14 pp, n.s.) understates the true causal effect of AI: the control benchmark itself has been pulled up by spillover or learning, so the HI "saturation" on the extensive margin is partly an artifact of a lifted counterfactual rather than evidence that AI has stopped working on conversion. The LO band, where within-broker spillover is mechanically smaller because each broker sees fewer treated customers, likely yields the cleaner ATE on the extensive margin: +1.13 pp ($p < 0.01$) on the 15-day LPM, almost twice the pooled estimate of +0.63 pp.

Read together, the band asymmetry and the control-group gap are consistent with a learning-by-doing story: brokers who accumulate more AI experience improve the *quality* of their deployment decisions, identifying which conversations and which suggestions are worth using, and the gains shift from expanding the pool of converting customers (extensive margin) toward raising per-buyer ticket size (intensive margin). They are also consistent with within-broker spillover: AI-suggested phrasing or framing learned in treated conversations bleeds into non-treated ones. The two channels are observationally equivalent in the data I have here; both lift the control benchmark for high-exposure brokers. Either way, ITT estimates from heavily-treated brokers should be read as a lower bound on AI's true contribution, because the counterfactual itself rises with cumulative AI exposure.

4.2 Robustness: middle-50% subsample

The learning result in Section 4.1 exploits cross-broker variation in treat share, raising the question of whether the pooled main results in Section 3 are driven by the tails. To check this, I restrict to the central 50 % of the treat-share distribution, $[0.757, 0.812]$, covering 539 brokers and 90,756 customers, where cross-broker variation in AI exposure is minimal. Table 10 reports the pooled and by-rank estimates on this trimmed subsample.

Outcome	Model	POOL	L1	L2	L3	L4+
Premium-15d	PPML	13.1	−43.8	6.0	35.0	16.4
		(11.9)	(26.6)	(16.6)	(28.9)	(24.0)
	Ext	0.52***	−0.98**	0.62***	0.51	0.89*
		(0.16)	(0.48)	(0.22)	(0.34)	(0.49)
	Ext-FE	0.37**	−0.97**	0.50**	0.19	0.78
		(0.16)	(0.48)	(0.21)	(0.32)	(0.51)
	IV-Ext	1.18***	−2.48*	1.46***	1.12	1.85*
		(0.37)	(1.36)	(0.51)	(0.75)	(0.96)
	Int log	12.5	−13.7	−5.9	54.9***	19.6
		(11.0)	(39.5)	(14.4)	(25.8)	(21.2)
Premium-1d	PPML	12.2	−56.0	7.0	31.8	19.8
		(13.4)	(25.0)	(19.2)	(31.9)	(26.9)
	Ext	0.50***	−0.85*	0.61***	0.56**	0.70
		(0.14)	(0.46)	(0.18)	(0.26)	(0.45)
	Ext-FE	0.37***	−0.82*	0.49***	0.29	0.58
		(0.14)	(0.44)	(0.18)	(0.25)	(0.46)
	IV-Ext	1.14***	−2.16*	1.42***	1.21*	1.45*
		(0.32)	(1.16)	(0.44)	(0.65)	(0.82)
	Int log	11.7	−49.3**	1.8	46.7*	20.1
		(13.0)	(21.4)	(18.3)	(32.2)	(27.7)
Dialogue rounds	PPML	6.5***	−0.3	7.3***	5.1**	9.0***
		(1.2)	(4.3)	(1.8)	(2.2)	(3.6)
	Int log	5.1***	2.8	5.2***	3.5**	7.7***
		(0.9)	(2.8)	(1.2)	(1.6)	(2.4)
N customers		90,756	6,527	48,404	21,595	14,230
N brokers		539	47	304	117	71

Table 10: Treatment effects on the middle-50% treat-share subsample, pooled and by broker rank

Notes: Sample restricted to brokers with treat share in $[0.757, 0.812]$ (p25–p75 of the cross-broker distribution). Each pair of rows reports the effect estimate (top) and standard error in parentheses (bottom). PPML and Int log effects in percent; Ext, Ext-FE, and IV-Ext in percentage points (pp). Standard errors clustered at the broker level. Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Pooled controls: customer-segment FE, first-day FE, broker-rank FE, broker-cohort FE. By-rank specifications drop broker-rank FE within each rank.

The pooled estimates survive the trimming with only modest attenuation: the IV-Extensive LATE on Premium-15d falls from +1.41 pp in the full sample (Table 6) to +1.18 pp here, and the within-broker LPM and broker-FE specifications retain 1 %-level significance. The intensive-margin PPML loses precision but keeps the positive

sign.

By rank, however, the full-sample monotone L1→L4+ extensive-margin gradient (Table 13) does *not* survive. L1 brokers (47 brokers, 6,527 customers) turn negative on every premium extensive-margin specification. L2 carries most of the pooled positive effect, reflecting that L2 makes up 56 % of subsample brokers. L3 and L4+ remain positive on the extensive margin but noisier than in the full sample. The dialogue-rounds gradient by rank survives cleanly across L1 to L4+.

The L1 reversal is consistent with the learning channel in Section 4.1: at moderate exposure, junior brokers have not accumulated enough experience to generate positive conversion effects, and may even generate negative ones when they routinely deploy AI suggestions that misfire in their hands. The full-sample positive L1 estimate appears to come disproportionately from the treat-heavy tail of L1 brokers, where higher exposure has either trained them up or produced more within-broker spillover. With only 47 brokers in the cell, this should be treated as suggestive rather than dispositive.

5 Customer Heterogeneity and Mechanisms

This section characterizes which customers benefit most from AI and what channels drive the conversion gain. I document that gains concentrate among lower-value and unfamiliar customers, and identify three mechanisms: AI sustains customer engagement rather than accelerating closing, operates as a discovery tool for new customers, and disproportionately raises high-value sales.

5.1 By customer segment

Table 11 reports ITT estimates on premium outcomes broken down by the customer's pre-experiment value segment.

Outcome	Model	B	A	S	S+	New
Premium-15d	PPML	45.35***	25.82*	28.97	27.07	22.76
		(19.9)	(16.4)	(38.6)	(28.3)	(29.7)
	Extensive	0.61***	0.68***	1.28*	−0.03	0.28
		(0.17)	(0.17)	(0.68)	(0.81)	(1.03)
	Ext Broker FE	0.38*	0.55***	1.02	−0.02	−0.20
		(0.17)	(0.17)	(0.71)	(0.85)	(1.09)
	IV-Extensive	1.57***	1.45***	2.54*	−0.07	0.50
		(0.45)	(0.37)	(1.38)	(1.56)	(1.80)
	Int log	22.46	8.67	−1.76	0.86	23.43
		(17.5)	(11.0)	(23.7)	(18.4)	(35.0)
Premium-1d	PPML	36.49*	29.29*	23.61	26.64	32.90
		(21.8)	(19.9)	(40.8)	(32.7)	(38.3)
	Extensive	0.61***	0.62***	0.50	−0.08	0.91
		(0.14)	(0.14)	(0.60)	(0.67)	(0.81)
	Ext Broker FE	0.41***	0.50***	0.13	−0.13	0.62
		(0.14)	(0.14)	(0.63)	(0.69)	(0.87)
	IV-Extensive	1.58***	1.33***	0.99	−0.15	1.66
		(0.38)	(0.31)	(1.21)	(1.32)	(1.58)
	Int log	17.07	18.87	16.14	−4.42	32.57
		(20.2)	(14.4)	(32.6)	(21.1)	(48.5)
N customers		61,571	81,871	8,078	8,727	4,411
N brokers		1,049	1,057	1,004	1,005	452

Table 11: Treatment effects on premium outcomes, by customer segment

Notes: Each pair of rows reports the effect estimate (top) and standard error in parentheses (bottom). PPML and Int log effects in percent; Extensive, Ext Broker FE, and IV-Extensive in percentage points (pp). Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Standard errors clustered at the broker level. Controls: first-day FE, broker-rank FE, broker-cohort FE (customer-segment FE omitted in stratum-specific regressions). Full regression coefficients ($\hat{\beta}$) and standard errors are reported in Tables A7 and A8 in Appendix C.

Effects decline with customer value. Conversion gains are largest at the bottom of the customer value scale. For segment B customers, the extensive-margin lift is +0.61 pp, corresponding to a +45 % PPML gain. Segment A produces a +0.68 pp extensive-margin effect and +26 % PPML. Segment S retains a marginally significant +1.28 pp extensive lift. Segment S+ produces estimates indistinguishable from zero across all models.

A natural interpretation is that high-value (S+) customers are already on a converting trajectory by the time they enter a brokered conversation, the broker has already established the relationship, and the customer arrives with strong prior intent. The AI tool therefore has little remaining headroom on which to operate for this group. The gains instead concentrate among mid- and low-value customers, for whom the advisory conversation is the primary persuasion and information channel.

Table 12 reports the corresponding effects on dialogue rounds.

Outcome	Model	B	A	S	S+	New
Dialogue rounds	PPML	10.31***	7.54***	14.32***	7.27	12.97*
		(2.09)	(2.08)	(5.30)	(5.21)	(7.97)
	Int log	5.31***	6.53***	8.53***	4.60	6.68*
		(0.96)	(0.93)	(3.05)	(2.97)	(4.05)
Pre-buy rounds	PPML	3.07	−11.62*	7.25	0.18	52.01***
		(8.6)	(6.7)	(12.1)	(11.1)	(23.2)
N customers		61,571	81,871	8,078	8,727	4,411
N brokers		1,049	1,057	1,004	1,005	452

Table 12: Treatment effects on dialogue rounds, by customer segment

Notes: Each pair of rows reports the effect estimate (top) and standard error in parentheses (bottom). Effects in percent. Pre-buy rounds defined on the purchaser subsample. Standard errors clustered at the broker level. Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Full regression coefficients ($\hat{\beta}$) and standard errors are reported in Table A11 in Appendix C.

New customers: a discovery channel. Customers without prior platform history display no detectable conversion gain but a striking +52 % ITT effect on Pre-buy rounds ($p < 0.01$). When a broker faces an unfamiliar customer, AI-enabled conversations run substantially longer before any purchase. This is consistent with the broker using AI suggestions to surface diagnostic questions about the customer's needs, budget, and family circumstances, questions that a junior broker or a broker unfamiliar with the customer might otherwise omit. The AI operates as a discovery tool for new customers, not as a closing accelerator.

Segment A is the only on-scale group with a marginally significant *negative* effect on Pre-buy rounds (−11.6 %, $p < 0.10$). For established mid-value customers, AI may help brokers close more efficiently, shortening the path to purchase for customers who were already engaged.

5.2 Engagement, not acceleration

The dialogue-rounds gain (+9 % PPML) combined with the near-zero pre-buy-rounds effect (−1.1 %, not significant) rules out "AI saves time by compressing the close" as the operative channel. Instead, AI raises conversion by sustaining conversations that would otherwise not reach the purchase stage. The treated customer engages in more total dialogue, but once they commit to purchase, the path from first contact to closing is the same length as in the control group.

A clean decomposition of this mechanism emerges from the heterogeneity in Section 6.1. Dialogue rounds rise monotonically with broker seniority (+3.8 % at L1 to +12.2 % at L4+), mirroring the conversion-rate gains. The engagement channel is not uniform: it is most powerful when the broker can use AI to diagnose customer needs and respond to follow-up questions with precision.

5.3 The discovery channel: new customers

The sharpest evidence on mechanism comes from the New customer segment, which has no prior platform history. For this group, the broker has no diagnostic baseline on which to build the conversation, making AI suggestions, which condition on customer-profile information, particularly valuable for the early stages of needs discovery.

The data confirm this interpretation. New customers show no detectable conversion gain (+0.28 pp, not significant) but a striking +52 % increase in pre-buy dialogue rounds ($p < 0.01$). AI-enabled conversations with new customers are substantially longer before any purchase occurs. This is consistent with AI helping brokers surface the diagnostic questions, about family composition, existing coverage, health history, and budget, that establish what the customer needs before a product recommendation is appropriate.

For segment A customers (established, mid-value), the pattern reverses: Pre-buy rounds is marginally significant and *negative* ($-11.6\,\%$, $p < 0.10$). For customers with a known profile, AI may instead help brokers organize and deliver the relevant information more efficiently, shortening the closing path. The two patterns, discovery for new customers, acceleration for established customers, are consistent with AI serving different functions at different stages of the customer relationship.

5.4 High-value sales

AI assignment expands the probability of a large-ticket sale (premium $\geq$ 200,000 RMB) by 0.067 pp ($p < 0.01$, LPM), representing a +160 % relative gain on a baseline probability of 0.04 %. The lower threshold of 100,000 RMB yields +0.156 pp ($p < 0.01$). The intensive-margin effect conditional on a large-ticket sale is +9.8 % (PPML) but not statistically significant.

This pattern is consistent with the engagement mechanism: AI sustains long, complex advisory conversations that would otherwise stall, enabling some customers to arrive at large-ticket product recommendations they would not have reached otherwise. Once a large-ticket sale occurs, its size is determined by the product itself rather than by the AI's assistance.

The large-ticket finding is also consistent with the amplification result: senior brokers, who benefit most from AI in overall terms, are also the brokers most likely to be matched with S+ customers and to handle the product complexity required for large-ticket policies.

6 Broker Heterogeneity and AI Adoption

This section asks which brokers gain most from AI and decomposes the answer into adoption frequency and per-use return. I document monotone amplification across broker seniority and show that an inverse adoption-frequency-versus-return pattern accounts for it: junior brokers adopt more often but extract less per use, while senior brokers adopt selectively and extract more.

6.1 By broker seniority: amplification

Table 13 reports ITT estimates on premium outcomes broken down by broker seniority.

Outcome	Model	L1	L2	L3	L4+
Premium-15d	PPML	47.47	30.62**	35.96*	24.28
		(78.3)	(14.2)	(21.5)	(20.5)
	Extensive	—	0.62***	0.83***	0.90**
		—	(0.16)	(0.25)	(0.37)
	Ext Broker FE	−0.37	0.43***	0.59**	0.59
		(0.37)	(0.16)	(0.24)	(0.38)
	IV-Extensive	−0.76	1.42***	1.83***	1.88**
		(0.96)	(0.37)	(0.57)	(0.79)
	Int log	−5.79	1.80	25.88*	25.06
		(28.6)	(9.93)	(17.1)	(18.1)
Premium-1d	PPML	45.37	32.75**	24.79	31.88
		(98.4)	(17.1)	(22.8)	(26.3)
	Extensive	—	0.61***	0.71***	0.78**
		—	(0.14)	(0.20)	(0.34)
	Ext Broker FE	−0.26	0.44***	0.50**	0.49
		(0.31)	(0.14)	(0.20)	(0.34)
	IV-Extensive	−0.59	1.40***	1.56***	1.63**
		(0.80)	(0.31)	(0.49)	(0.68)
	Int log	−25.10	10.09	31.66	22.96
		(23.9)	(12.7)	(22.6)	(22.5)
N customers		13,335	91,682	37,798	21,843
N brokers		132	594	217	119

Table 13: Treatment effects on premium outcomes, by broker seniority

Notes: Each pair of rows reports the effect estimate (top) and standard error in parentheses (bottom). PPML and Int log in percent; Extensive, Ext Broker FE, and IV-Extensive in percentage points (pp). The L1 subsample contains only 132 brokers; cluster-robust standard errors fail to converge in some cells ("—"), and L1 point estimates should be treated as suggestive. Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Standard errors clustered at the broker level. Controls: customer-segment FE, first-day FE, broker-cohort FE (broker-rank FE omitted in rank-specific regressions). Full regression coefficients ($\hat{\beta}$) and standard errors are reported in Tables A9 and A10 in Appendix C.

Monotone amplification. Effects rise monotonically with broker seniority on every margin. On the extensive margin, the fifteen-day conversion lift is +0.62 pp at L2 (the smallest group with reliable inference), rising to +0.83 pp at L3 and +0.90 pp at L4+. The most senior brokers' LATE on actually-AI-using conversations reaches +1.88 pp.

The PPML gains on the premium level are similarly large across L2 through L4+ (e.g., +30.6 % at L2 and +36.0 % at L3, with L4+ comparable; see Table 13). On dialogue volume (Table 14), Dialogue rounds rises from +3.8 % (L1, not significant) to +12.2 % (L4+, $p < 0.01$).

Outcome	Model	L1	L2	L3	L4+
Dialogue rounds	PPML	3.80	8.33***	10.43***	12.20***
		(5.00)	(1.93)	(2.43)	(4.03)
	Int log	—	5.87***	5.33***	8.37***
		—	(0.90)	(1.28)	(1.94)
Pre-buy rounds	PPML	−29.93	−3.37	11.39	4.95
		(15.9)	(6.5)	(8.85)	(10.7)
N customers		13,335	91,682	37,798	21,843
N brokers		132	594	217	119

Table 14: Treatment effects on dialogue rounds, by broker seniority

Notes: Each pair of rows reports the effect estimate (top) and standard error in parentheses (bottom). Effects in percent. Pre-buy rounds defined on the purchaser subsample. Standard errors clustered at the broker level. L1 Int log cluster SE does not converge ("—"). Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Full regression coefficients ($\hat{\beta}$) and standard errors are reported in Table A12 in Appendix C.

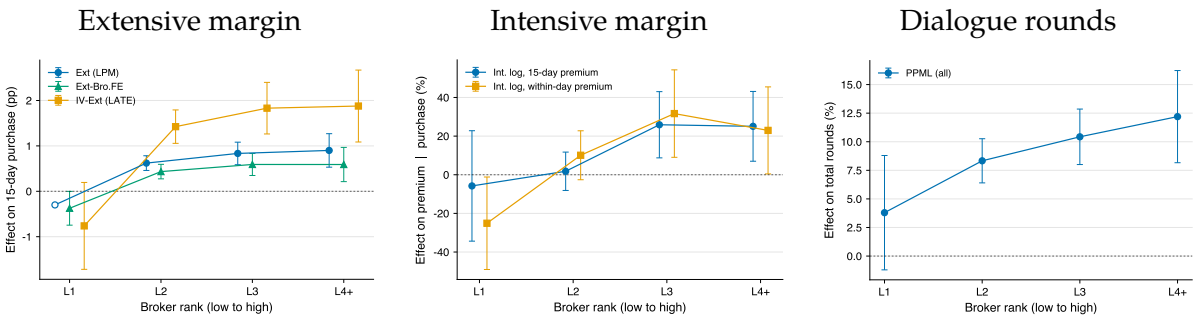


Figure 5: ITT estimates by broker seniority rank

Notes: Each panel plots the treatment effect estimate (point estimate and 95 % confidence interval) for brokers at each seniority level. *Left:* extensive-margin LPM in pp (15-day window). *Center:* intensive-margin Int log in % (conditional on purchase). *Right:* Dialogue rounds PPML in %. Bars use broker-clustered standard errors. The L1 estimates carry wide confidence intervals reflecting the small number of brokers (132) at that level.

The L1 subsample (132 brokers) is too small for precise inference; cluster-robust standard errors fail to converge in some cells. The point estimates are near zero or negative for L1, but should be treated as suggestive. The gradient from L2 through L4+ is precisely estimated and monotone.

AI as a complement to broker human capital. The monotone amplification pattern is inconsistent with AI functioning as a substitute for broker skill, which would predict larger gains for less-skilled (junior) brokers, as in Brynjolfsson et al. (2025). Instead,

senior brokers extract larger gains from the same set of AI suggestions, presumably because they bring more contextual judgment to the question of when a suggestion is appropriate and how to integrate it with the customer's specific circumstances. This finding qualifies the common characterization of AI as an equalizer in professional labor markets: in a high-complexity advisory setting, AI amplifies existing skill differences.

6.2 Adoption frequency vs. per-adoption return

Figures 6 and 7 document the joint pattern of adoption behavior and realized gains across broker seniority levels.

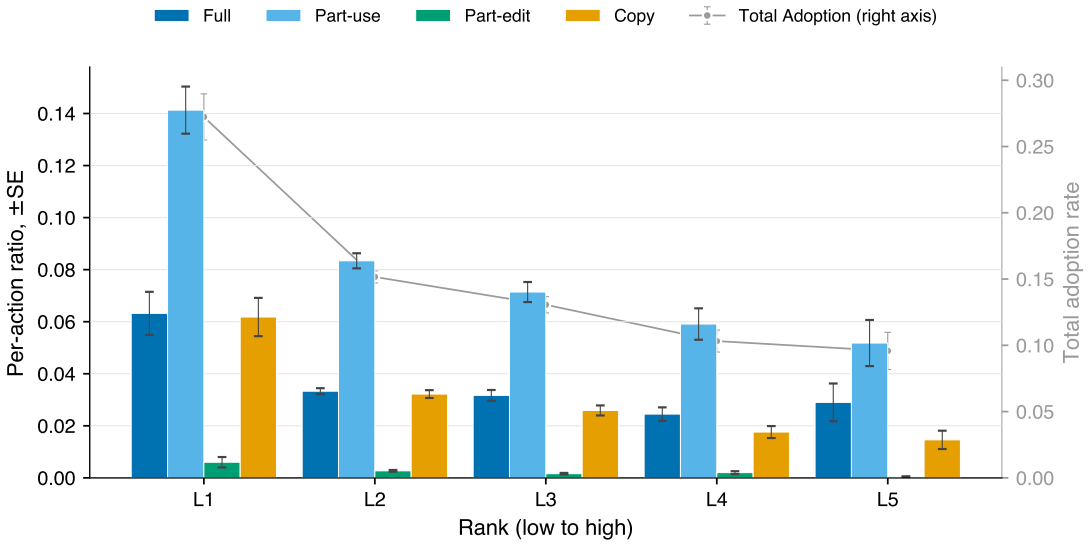


Figure 6: Aggregate AI-suggestion handling rates by broker seniority

Notes: Each bar shows the mean broker-level active adoption rate (AI suggestions acted upon in any of the four handling categories) for a given seniority rank. Total adoption averages approximately 17 % across all brokers; the 95th percentile reaches 40 % and the 99th percentile 64 %. Active adoption falls monotonically from L1 (most junior) to L5 (most senior), consistent with the broker-level a_b gradient discussed below in Section 6.1.

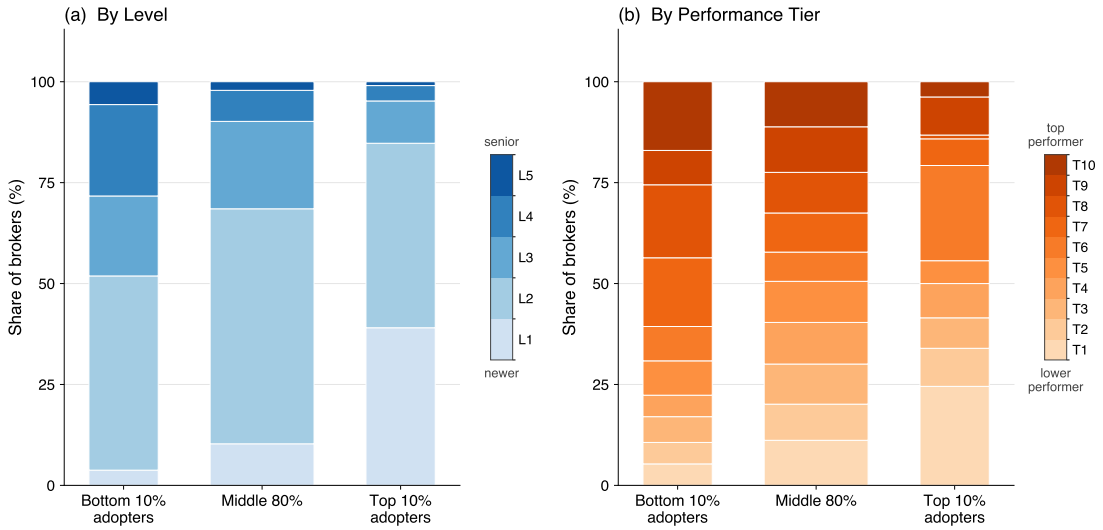


Figure 7: Broker characteristics: top vs. bottom AI adopters

Notes: Comparison of observable characteristics for the top quartile (highest active-adoption rate) vs. bottom quartile (lowest active-adoption rate) of brokers in the treatment group. Top adopters are disproportionately junior, shorter-tenured, and from lower seniority ranks; bottom adopters are disproportionately senior with longer tenure and higher seniority. This confirms that the adoption-frequency gradient is not driven by individual differences in effort or engagement, but by systematic differences in how senior vs. junior brokers relate to the AI tool.

The figures deliver a clear message. Junior brokers adopt a higher share of AI suggestions outright. Senior brokers adopt less frequently but, when they do, they are more likely to edit the suggestion, a pattern consistent with the broker exercising judgment about whether and how to use the AI's output. The brokers who use AI most often are not the brokers who derive the largest revenue gains from it.

Let a_b denote a broker's per-suggestion adoption rate (the share of AI suggestions she acts on, summed across the four handling categories) and LATE_b the per-use return within her caseload. The ITT–LATE decomposition in Section 3 then provides the accounting: junior brokers have higher a_b but lower LATE_b; senior brokers have lower a_b but substantially higher LATE_b. The per-adoption return gradient dominates, producing the amplification result. This implies that adoption-rate metrics alone, which firms commonly use to track AI rollouts, overstate the realized value created for junior brokers and understate it for senior brokers.

7 Conclusion

This paper presents experimental evidence on the productivity effects of a production-grade generative-AI tool deployed to licensed insurance brokers at Ant Insurance, the largest insurance brokerage in China. Four findings stand out.

First, generative AI substantially raises realized revenue. The fifteen-day conversion probability rises by 0.63 percentage points (+16 % relative), and total premium per cus-

tomer rises by approximately +30 % (PPML). The effect operates primarily through the extensive margin, more customers reach a purchase decision, rather than by increasing the per-transaction premium of customers who were already converting. The AI sustains engagement volume (+9 % dialogue rounds) without shortening the closing path for already-converting customers.

Second, AI amplifies existing productivity differences across brokers rather than leveling them. Intent-to-treat effects rise monotonically with broker seniority, from near zero for the most junior brokers to 0.90 percentage points for the most senior. This pattern departs from the skill-leveling finding documented in the prior literature on AI in low-stakes white-collar tasks. The interpretation is that effectively using AI in high-complexity advisory requires contextual judgment, knowing when a suggestion is appropriate, and that judgment is strongly correlated with broker experience.

Third, brokers learn from cumulative AI exposure. Comparing brokers in the high- and low-exposure tails of the treat-share distribution, AI experience shifts gains from the extensive margin (in low-exposure brokers) toward the intensive margin (in high-exposure brokers), and lifts conversion rates among control customers served by high-exposure brokers, indicating broker upskilling that transfers to non-AI interactions and that is observationally consistent with within-broker spillover.

Fourth, adoption frequency and per-adoption return are inversely related across broker ranks. Junior brokers adopt AI suggestions more often; senior brokers adopt less often but extract larger gains per adoption. Firms that track AI rollouts by adoption-rate metrics will therefore overstate the realized value created for junior brokers and understate it for senior brokers.

Several considerations bear on the external validity of these findings. First, the experiment runs on a single platform in the Chinese market over a twenty-day window; the estimates therefore capture the intent-to-treat effect of a contemporaneous rollout, not the long-run equilibrium that would emerge after brokers, customers, and the platform have fully adapted to the AI tool. Second, whether the amplification result generalizes to other professional services, medicine, law, financial planning, depends on whether those settings share three key structural features of the present one: high dispersion in worker ability, information-intensive tasks that require complex customization, and high per-decision stakes. These features are common in professional advisory roles, suggesting that the qualitative pattern may extend beyond insurance brokerage, though specific magnitudes will differ.

Buyer-side welfare is not measured here. If the AI tool facilitates product-mix shifts toward higher-premium, higher-commission policies, the seller-side revenue gain need not translate into a buyer-side welfare gain. Answering this question requires richer customer-side outcomes, which I plan to obtain in the next phase of this project.

The findings of this paper suggest several implications for AI deployment in profes-

sional services. A uniform rollout will not deliver inclusive productivity gains when effects amplify at the top of the skill distribution; platforms and firms should consider conditioning deployment on broker seniority and task complexity, and investing in complementary training for junior brokers to raise their per-adoption return. More broadly, the inverse adoption-frequency-versus-return pattern implies that deployment decisions informed only by adoption-rate metrics will systematically misallocate AI resources toward workers for whom the per-adoption return is lowest.

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.

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*, 2025. Advance access, DOI 10.1093/restud/rdaf084; NBER Working Paper 31747.

Augenblick, Ned, Eben Lazarus, and Michael Thaler, "Overinference from Weak Signals and Underinference from Strong Signals," *Quarterly Journal of Economics*, 2025, *140* (1), 335–401.

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

Baslandze, Salomé, Zachary Edwards, John R. Graham, Ty McClure, Michael Sparks, Brent Meyer, Sonya Ravindranath Waddell, and Daniel Weitz, "Artificial Intelligence, Productivity, and the Workforce: Evidence from Corporate Executives," *Federal Reserve Banks of Atlanta and Richmond Working Paper*, 2026.

Bastani, Hamsa, Osbert Bastani, and Wichinpong P. Sinchaisri, "Improving Human Sequential Decision Making with Reinforcement Learning," *Management Science*, 2025, *72* (1), 733–755.

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, Utpal, Amit Kumar, Sujata Visaria, and Jing Zhao, "Do Women Receive Worse Financial Advice?," *Journal of Finance*, 2024, *79* (5), 3261–3307.

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

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts, "Does Management Matter? Evidence from India," *Quarterly Journal of Economics*, 2013, *128* (1), 1–51.

Brynjolfsson, Erik, 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.

Bucher-Koenen, Tabea, Andreas Hackethal, Johannes Koenen, and Christine Laudenbach, "Gender Differences in Financial Advice," *American Economic Review*, 2025, *115* (12), 4218–4252.

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," *Working Paper*, 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. Forthcoming, DOI 10.1287/mnsc.2025.00535.

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.

de Souza, Gustavo, "AI in the Office and the Factory: Evidence from Administrative Software Registry Data," *Federal Reserve Bank of Chicago Working Paper 2025-11*, 2026.

Dell'Acqua, Fabrizio, "Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters," *Working Paper*, 2022.

_ , Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *Harvard Business School Working Paper 24-013*, 2023.

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.

Egan, Mark L., Gregor Matvos, and Amit Seru, "Arbitration with Uninformed Consumers," *Review of Economic Studies*, 2025, *92* (6), 3888–3923.

Egan, Mark, 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.

Emanuel, Natalia, Emma Harrington, and Amanda Pallais, "The Power of Proximity to Coworkers: Training for Tomorrow or Productivity Today?," *NBER Working Paper 31880*, 2023. Conditionally accepted, Quarterly Journal of Economics.

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

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

Guiso, Luigi, Andrea Pozzi, Anton Tsoy, Leonardo Gambacorta, and Paolo Emilio Mistrulli, "The Cost of Steering in Financial Markets: Evidence from the Mortgage Market," *Journal of Financial Economics*, 2022, *143* (2), 1209–1226.

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

Huang, Shawn, Artur Hugon, Tengfei Zhang, and Wenting Zheng, "Generative AI and Investment Research: Evidence from Analyst Reports," *Working Paper*, 2025.

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

Ibrahim, Rouba, Sang-Hyun Kim, and Jordan Tong, "Eliciting Human Judgment for Prediction Algorithms," *Management Science*, 2021, *67* (4), 2314–2325.

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

Jaffe, Sonia, Neha Parikh Shah, Jenna Butler, Alex Farach, Alexia Cambon, Brent Hecht, Michael Schwarz, and Jaime Teevan, "Generative AI in Real-World Workplaces," *Microsoft Research*, 2024.

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.

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 Information Production in Selection Markets: Experimental Evidence from Insurance Intermediation," *Working Paper*, 2024.

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.

McLaughlin, Bryce and Jann Spiess, "Designing Algorithmic Recommendations to Achieve Human-AI Complementarity," *arXiv:2405.01484*, 2024.

Mozannar, Hussein and David Sontag, "Consistent Estimators for Learning to Defer to an Expert," in "International Conference on Machine Learning (ICML)" 2020, pp. 7076–7087.

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 Philip Clarke, Solène Delecourt, David Holtz, and Rembrand Koning, "The Uneven Impact of Generative AI on Entrepreneurial Performance," *Harvard Business School Working Paper 24-042*, 2024.

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.

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan, "The Algorithmic Automation Problem: Prediction, Triage, and Human Effort," *arXiv:1903.12220*, 2019.

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.

Skitka, Linda J., Kathleen L. Mosier, and Mark D. Burdick, "Does Automation Bias Decision-Making?," *International Journal of Human-Computer Studies*, 1999, *51*, 991–1006.

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*, 2025. Forthcoming.

A Experimental Details, Balance Checks, and Broker Adoption

This appendix collects supporting material for the experimental setup described in Section 2. I first show the platform interfaces through which customers and brokers interact and the broker workstation in which the AI tool is deployed. I then document sample composition and balance across the treatment and control groups: customer age, customer-value tiers, and the cross-broker distribution of treat share. The final subsection examines how the broker-level AI adoption rate varies with seniority, which underlies the ITT $\approx \hat{\pi} \times$ LATE decomposition in the main text.

Platform interfaces and the AI workstation

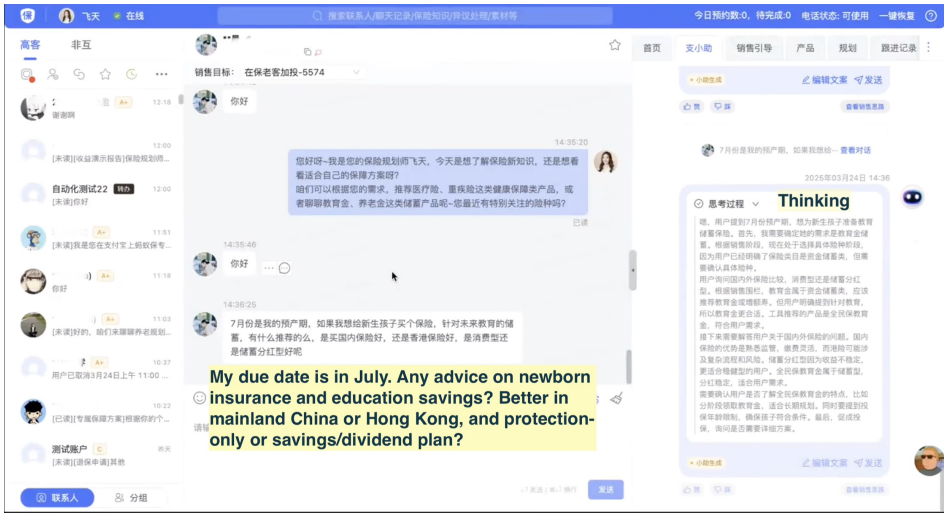
Figure A1 shows the customer-facing screens through which customers initiate contact with a broker. Figure A2 shows the broker's workstation with the AI tool active.



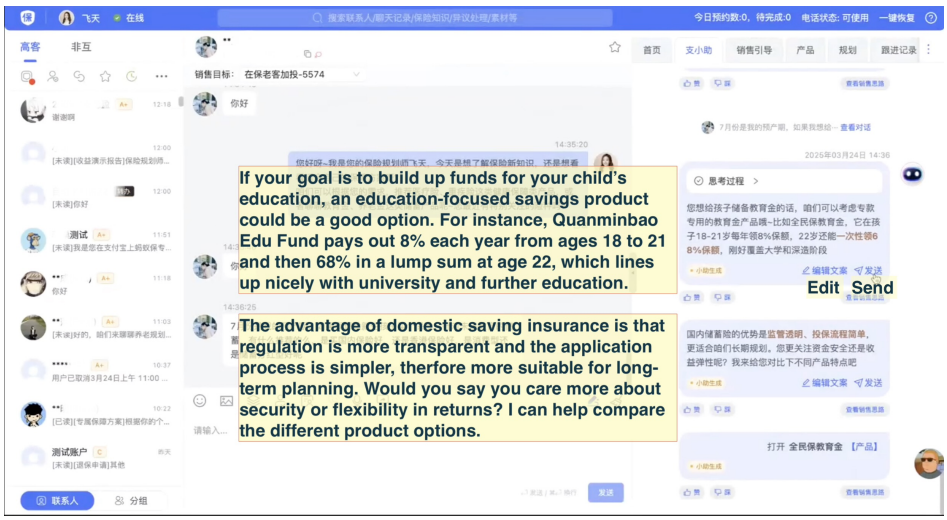
(a) Product page (b) IM chat interface (c) Broker profile (d) My insurance

Figure A1: Customer-facing interface: how a customer meets a broker

Notes: Screenshots from the Alipay app showing the four main screens through which a customer engages with a broker. (a) A product detail page listing coverage terms, premiums, and an option to connect with a licensed broker for advice. (b) The instant-messaging (IM) chat interface, where the advisory conversation takes place. This is the channel in which the AI tool is active for treated conversations; AI suggestions reach the customer only if the broker chooses to send them. (c) The broker's public profile page, displaying seniority rank, customer reviews, and areas of expertise. (d) The customer's personal insurance portfolio page, summarizing existing policies. All text in the screenshots is in Chinese; no customer data is displayed.



(a) AI processes the customer message



(b) Reply suggestions appear in the broker's panel

Figure A2: Broker workstation with the AI tool active

Notes: Screenshots from the broker's workspace. (a) Upon receiving a customer IM message, the AI tool (*Tongtiandai*) processes it in real time, drawing on the fine-tuned Qwen2.5-72B model, the insurance knowledge base, and past advisory transcripts. (b) The tool returns a ranked list of suggested replies in a side panel visible only to the broker; the customer sees none of this. The broker reviews the suggestions and decides whether to adopt, modify, or discard each one, producing the four handling-action categories analyzed below.

Sample composition and balance

I document balance across the treatment and control groups along three dimensions: customer age (Figure A3), the cross-broker distribution of treat share (Table A1), and the customer distribution and conversion rates by value tier (Table A2).

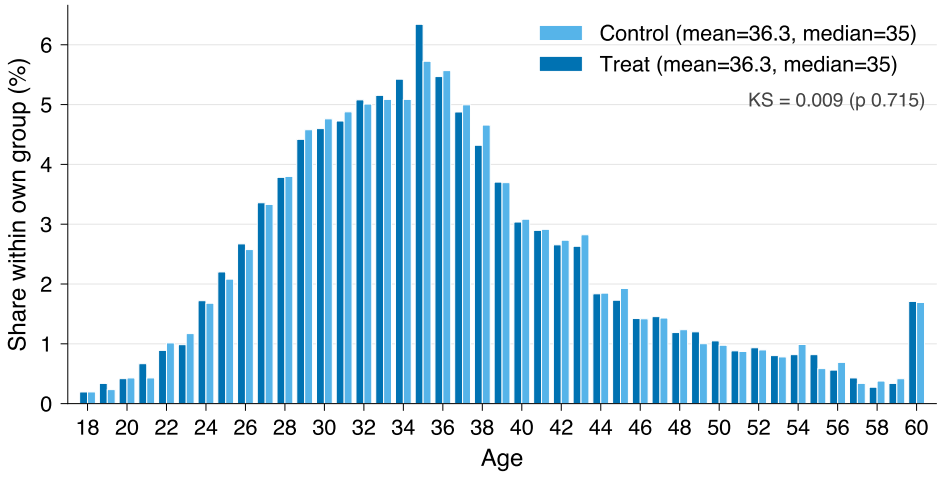


Figure A3: Age distribution of customers, by treatment group

Notes: Kernel density estimates of customer age for the treatment group (AI enabled, n = 128,957) and the control group (n = 35,701). The two distributions are nearly identical, confirming that the customer-level randomization produced balanced treatment and control groups on age. Customer age is reported as of the date of first contact with a broker. The customer base is predominantly working-age adults with the modal age in the mid-thirties.

The cross-broker distribution of treat share is tightly concentrated around the 78 % randomization target. The tails used to define the low- and high-exposure bands in Section 4 (below p25 and above p75) are driven by small caseload sizes rather than by any systematic broker characteristic; 89.3 % of brokers fall within [0.70, 0.85].

n_{brk}	mean	sd	min	p1	p5	p25	p50	p75	p95	p99	max
1,071	0.780	0.083	0.000	0.500	0.702	0.757	0.784	0.812	0.857	1.000	1.000

Table A1: Cross-broker distribution of per-broker treat share

Notes: 1,071 brokers in the engaged-customer sample; 164,658 customers. Treat share = #treated customers/#customers per broker. The five brokers with treat share = 0 and the 24 with treat share = 1 have very small caseloads.

The customer-level distribution by value tier and the corresponding 15-day conversion rates provide the denominators for the heterogeneity estimates in Section 5. Conversion rates rise monotonically from 3.28 % (B) to 10.04 % (S+), reflecting the platform's value-segment ordering. The "New" tier's relatively high conversion rate (8.52 %) reflects that these customers have actively sought advice, even without a prior purchase history. Within-day conversion rates (Y1) mirror the 15-day pattern: B 2.34 %, A 3.09 %, S 5.29 %, S+ 6.86 %, New 6.41 %.

Tier	Total customers	Within-day buyers	15-day buyers	15-day conversion
B	61,571	1,441	2,017	3.28%
A	81,871	2,530	3,450	4.21%
S	8,078	427	571	7.07%
S+	8,727	599	876	10.04%
New	4,411	283	376	8.52%
Total	164,658	5,280	7,291	4.43%

Table A2: Customer counts and 15-day conversion rates by tier

Notes: "15-day buyers" are customers with strictly positive 15-day premium. "15-day conversion" is the ratio of 15-day buyers to total customers. The baseline rates contextualize the heterogeneous treatment effects in Tables 11 and 12: a +0.83 pp extensive-margin ITT for S customers represents a 12 % relative gain on a 7.07 % base, while an insignificant −0.03 pp for S+ operates against a 10.04 % base that leaves less room for AI to expand the marginal converter pool.

Broker AI-suggestion adoption by seniority

The main text reports a pooled active-adoption rate of 44.3 % for treated, engaged customers and uses this figure to anchor the decomposition ITT $\approx \hat{\pi} \times$ LATE. This subsection documents how adoption behavior varies across broker seniority ranks. The platform records four handling actions for every AI suggestion exposed to a treated broker: (i) fully adopt and send with one click, (ii) partially adopt after deleting part of the suggestion, (iii) partially adopt after editing, and (iv) copy as inspiration without sending in this conversation. Table A3 reports the cross-broker mean of the broker-level ratio #action/#AI suggestions exposed, with each broker weighted equally within its seniority rank. "Active adoption" sums all four action categories.

Action	L1	L2	L3	L4	L5
N brokers	132	591	215	93	25
Adopt fully (one-click send)	6.32	3.33	3.17	2.45	2.90
Adopt partially (send after deleting)	14.13	8.34	7.14	5.91	5.18
Adopt partially (send after editing)	0.60	0.27	0.16	0.20	0.04
Copy as inspiration	6.18	3.22	2.59	1.76	1.46
Active adoption (sum of all four)	**27.23**	**15.16**	**13.06**	**10.32**	**9.58**

Table A3: Broker-average AI-suggestion handling rates, by seniority (%)

Notes: Each cell is the cross-broker mean of the broker-level ratio #action/#AI suggestions exposed within the seniority rank, with each broker weighted equally. Action categories are not mutually exclusive (a single AI suggestion can trigger more than one), so within-broker shares can sum to slightly above 100 %. Mapping to the platform's internal labels: L1 = B0, L2 = B1, L3 = B2, L4 = B3, L5 = B4.

Two patterns emerge. First, active adoption declines monotonically with seniority: L1 brokers adopt roughly 2× as often as L2 brokers and 3× as often as L4–L5 brokers. This is the broker-level adoption-frequency gradient a_b that maps to the per-use return

decomposition discussed in Section 6.1. Second, the decline is driven by all four action categories: partial sends, partial edits, and copy-as-inspiration all fall with seniority alongside full adoption. Senior brokers are therefore more selective, not merely less engaged: they adopt less frequently but, as shown by the ITT estimates in Section 6.1, extract a higher per-adoption return when they do use the tool.

B Customer Engagement Analysis

The main-text pre-buy rounds outcome (*PreBuyRounds*) is defined for same-day buyers only (customers with a positive within-day premium, $n = 5{,}280$): cumulative dialogue rounds up to and including the first purchase day. This discards 2,011 customers who purchase through the 15-day window without ever recording a same-day premium. This appendix extends the definition to all 15-day buyers and shows that the near-zero pooled pre-buy-rounds effect in the main text conceals a significant positive effect for the 15-day-only subgroup.

Extended definition. For all 15-day buyers, define

$$\text{PreBuyRounds}^{\text{ext}} = \begin{cases} \sum_{t \le t^*_{\text{prem1}>0}} \text{rounds}_t & \text{if prem1} > 0 \text{ (same-day buyer)} \\ \sum_{t \in \text{sample}} \text{rounds}_t & \text{if prem15} > 0 \text{ and prem1} = 0 \text{ (15-day-only buyer)} \end{cases}$$

For 15-day-only buyers, there is no recorded same-day purchase timestamp, so the full 20-day sample window is used as the upper bound on pre-buy rounds.

Quantity	Value
N ever-talked customers	164,658
N with 15-day premium > 0	7,291
of which same-day buyers (prem1>0)	5,280
of which 15-day-only buyers (prem15> 0, prem1=0)	2,011
Mean $\text{PreBuyRounds}^{\text{ext}}$ | same-day buyer	18.0
Mean $\text{PreBuyRounds}^{\text{ext}}$ | 15-day-only buyer	8.8

Table A4: Sample composition and mean pre-buy chat rounds, extended definition

The 15-day-only buyers have a substantially shorter average pre-purchase conversation (8.8 vs. 18.0 rounds), consistent with their purchase not being concentrated on a single high-intensity day.

	All 15-day buyers (prem15>0)	Same-day buyers (prem1>0)	15-day-only buyers (prem15>0, prem1=0)
Panel A: PPML (%)			
Effect	+2.46	−1.07	+10.76
	(4.24)	(4.59)	(9.06)
Panel B: Int log (%)			
Effect	+5.28*	−0.02	+13.10***
	(3.18)	(3.62)	(5.23)
N customers (PPML)	7,291	5,280	2,011
N customers (Int log)	7,239	5,228	2,011
N brokers	995	975	830

Table A5: Treatment effect on pre-buy chat rounds, extended definition

Notes: Effects in percent. PPML estimated on the level of $\text{PreBuyRounds}^{\text{ext}}$ (zeros included); Int log is OLS on $\log(\text{PreBuyRounds}^{\text{ext}})$ over the positive subsample. Controls: customer-segment FE, first-day FE, broker-rank FE, broker-cohort FE. Standard errors clustered at the broker level. Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

The pooled pre-buy rounds effect is driven entirely by the 15-day-only subgroup: the same-day sample (which replicates the main-text specification) shows a near-zero coefficient (−1.1 % PPML, −0.02 % Int log), matching the main-text result, while the 15-day-only subgroup shows +13.1 % (Int log, $p < 0.01$). This is consistent with AI helping brokers sustain longer conversations with customers whose purchase materializes later in the 15-day window.

Table A6 shows that the 15-day-only effect concentrates in higher-value segments. Among 15-day-only buyers, AI raises pre-buy chat rounds substantially in the B segment (+28 % PPML, $p < 0.10$), the A segment (+12.7 % Int log, $p < 0.10$), and especially S+ (+60 % PPML, $p < 0.01$; +30 % Int log, $p < 0.05$), precisely the customer cohort whose purchase decision takes time to materialize. This reinforces the discovery-channel interpretation in the main text: AI is most useful when conversations need to cover more ground before a purchase recommendation is appropriate.

	B	A	S	S+	New
Panel A: All 15-day buyers (prem15>0)					
PPML %	+10.25	−8.34	+4.49	+11.65	+45.59***
	(8.16)	(6.23)	(10.98)	(10.73)	(19.86)
Int log %	+5.91	+2.12	−14.54	+14.83*	+26.92**
	(6.09)	(4.75)	(8.24)	(9.33)	(14.95)
Panel B: Same-day buyers (prem1>0)					
PPML %	+2.80	−11.60*	+7.25	+0.24	+52.01***
	(8.49)	(6.73)	(12.14)	(11.19)	(23.24)
Int log %	−3.39	−3.78	−10.52	+6.74	+29.24**
	(6.81)	(5.29)	(10.05)	(10.90)	(16.71)
Panel C: 15-day-only buyers (prem15>0, prem1=0)					
PPML %	+27.90*	−1.73	−18.83	+59.74***	+7.51
	(17.16)	(13.91)	(19.19)	(24.97)	(24.54)
Int log %	+15.93	+12.67*	−7.60	+29.67**	+18.21
	(9.41)	(8.06)	(21.33)	(16.80)	(25.33)

Table A6: Treatment effect on pre-buy chat rounds, by customer segment and buyer type
Notes: Effects in percent. Standard errors clustered at the broker level. Controls: first-day FE, broker-rank FE, broker-cohort FE (customer-segment FE omitted within segment). Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Sample sizes by panel: A: B 2,018/2,004, A 3,451/3,433, S 571/567, S+ 876/866, New 376/369. B: B 1,442/1,428, A 2,531/2,513, S 427/423, S+ 597/588, New 283/283. C: B 576, A 568, S 144, S+ 278, New 93.

C Additional Tables on Regression Results

This appendix provides full regression output for the heterogeneity tables in the main text. Each table reports raw regression coefficients ($\hat{\beta}$) alongside the economically interpretable effect (in percent or percentage points), with broker-clustered standard errors. The pooled-sample results are already reported in full in Tables 6 and 7 of the main text.

C.1 Premium outcomes, by customer segment

	Pooled	B	A	S	S+	New
Panel A: PPML (%, level effect)						
$\hat{\beta}$	0.268	0.374	0.230	0.254	0.240	0.205
	(0.078)	(0.137)	(0.130)	(0.299)	(0.223)	(0.242)
Effect (%)	30.72***	45.35***	25.82*	28.97	27.07	22.76
	(10.20)	(19.87)	(16.37)	(38.61)	(28.31)	(29.70)
Panel B: Extensive (LPM, pp)						
$\hat{\beta}$	0.006	0.006	0.007	0.013	0.000	0.003
	(0.001)	(0.002)	(0.002)	(0.007)	(0.008)	(0.011)
Effect (pp)	0.63***	0.61***	0.68***	1.28*	−0.03	0.28
	(0.121)	(0.169)	(0.171)	(0.677)	(0.812)	(1.026)
Panel C: Extensive Broker FE (within-broker LPM, pp)						
$\hat{\beta}$	0.004	0.004	0.006	0.010	0.000	−0.002
	(0.001)	(0.002)	(0.002)	(0.007)	(0.008)	(0.011)
Effect (pp)	0.43***	0.38*	0.55***	1.02	−0.02	−0.20
	(0.120)	(0.170)	(0.166)	(0.710)	(0.849)	(1.087)
Panel D: IV-Extensive (2SLS, pp)						
$\hat{\beta}$	0.014	0.016	0.014	0.025	−0.001	0.005
	(0.003)	(0.004)	(0.004)	(0.014)	(0.016)	(0.018)
Effect (pp)	1.41***	1.57***	1.45***	2.54*	−0.07	0.50
	(0.276)	(0.449)	(0.365)	(1.379)	(1.562)	(1.802)
Panel E: Int log (intensive log OLS, %)						
$\hat{\beta}$	0.115	0.203	0.083	−0.018	0.009	0.211
	(0.069)	(0.143)	(0.101)	(0.242)	(0.182)	(0.284)
Effect (%)	12.21*	22.46	8.67	−1.76	0.86	23.43
	(7.75)	(17.48)	(11.03)	(23.74)	(18.38)	(35.04)
N customers	164,658	61,571	81,871	8,078	8,727	4,411
N brokers	1,063	1,049	1,057	1,004	1,005	452
Segment FE	Y	–	–	–	–	–
First-day FE	Y	Y	Y	Y	Y	Y
Broker-rank FE	Y	Y	Y	Y	Y	Y
Broker-cohort FE	Y	Y	Y	Y	Y	Y

Table A7: Full regression output: Premium-15d, by customer segment

Notes: Raw regression coefficients ($\hat{\beta}$) with broker-clustered standard errors in parentheses. Effects: % for PPML and Int log ($100(e^{\hat{\beta}}-1)$); pp for Extensive, Extensive Broker FE, and IV-Extensive ($100\hat{\beta}$). Int log estimated on the purchaser subsample: N = 7,290 pooled, 2,017 (B), 3,450 (A), 571 (S), 876 (S+), 376 (New). Extensive Broker FE drops single-group brokers, reducing N by 0.05–14 % (largest reduction in S and S+). Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

	Pooled	B	A	S	S+	New
Panel A: PPML (%, level effect)						
$\hat{\beta}$	0.264	0.311	0.257	0.212	0.236	0.284
	(0.092)	(0.160)	(0.154)	(0.330)	(0.258)	(0.288)
Effect (%)	30.23***	36.49*	29.29*	23.61	26.64	32.90
	(12.01)	(21.77)	(19.89)	(40.79)	(32.66)	(38.27)
Panel B: Extensive (LPM, pp)						
$\hat{\beta}$	0.006	0.006	0.006	0.005	−0.001	0.009
	(0.001)	(0.001)	(0.001)	(0.006)	(0.007)	(0.008)
Effect (pp)	0.58***	0.61***	0.62***	0.50	−0.08	0.91
	(0.103)	(0.140)	(0.143)	(0.601)	(0.665)	(0.813)
Panel C: Extensive Broker FE (within-broker LPM, pp)						
$\hat{\beta}$	0.004	0.004	0.005	0.001	−0.001	0.006
	(0.001)	(0.001)	(0.001)	(0.006)	(0.007)	(0.009)
Effect (pp)	0.41***	0.41***	0.50***	0.13	−0.13	0.62
	(0.102)	(0.141)	(0.141)	(0.628)	(0.690)	(0.865)
Panel D: IV-Extensive (2SLS, pp)						
$\hat{\beta}$	0.013	0.016	0.013	0.010	−0.002	0.017
	(0.002)	(0.004)	(0.003)	(0.012)	(0.013)	(0.016)
Effect (pp)	1.32***	1.58***	1.33***	0.99	−0.15	1.66
	(0.237)	(0.381)	(0.314)	(1.207)	(1.315)	(1.578)
Panel E: Int log (intensive log OLS, %)						
$\hat{\beta}$	0.154	0.158	0.173	0.150	−0.045	0.282
	(0.082)	(0.172)	(0.121)	(0.281)	(0.220)	(0.366)
Effect (%)	16.68*	17.07	18.87	16.14	−4.42	32.57
	(9.63)	(20.19)	(14.41)	(32.58)	(21.07)	(48.50)
N customers	164,658	61,571	81,871	8,078	8,727	4,411
N brokers	1,063	1,049	1,057	1,004	1,005	452
Segment FE	Y	–	–	–	–	–
First-day FE	Y	Y	Y	Y	Y	Y
Broker-rank FE	Y	Y	Y	Y	Y	Y
Broker-cohort FE	Y	Y	Y	Y	Y	Y

Table A8: Full regression output: Premium-1d, by customer segment

Notes: Raw regression coefficients ($\hat{\beta}$) with broker-clustered standard errors in parentheses. Effects: % for PPML and Int log ($100(e^{\hat{\beta}}-1)$); pp for Extensive, Extensive Broker FE, and IV-Extensive ($100\hat{\beta}$). Int log estimated on the same-day purchaser subsample: N = 5,280 pooled, 1,441 (B), 2,530 (A), 427 (S), 599 (S+), 283 (New). Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

C.2 Premium outcomes, by broker seniority

	Pooled	L1	L2	L3	L4+
Panel A: PPML (%, level effect)					
$\hat{\beta}$	0.268	0.388	0.267	0.307	0.217
	(0.078)	(0.531)	(0.109)	(0.158)	(0.165)
Effect (%)	30.72***	47.47	30.62**	35.96*	24.28
	(10.20)	(78.27)	(14.24)	(21.47)	(20.55)
Panel B: Extensive (LPM, pp)					
$\hat{\beta}$	0.006	—	0.006	0.008	0.009
	(0.001)	—	(0.002)	(0.002)	(0.004)
Effect (pp)	0.63***	—	0.62***	0.83***	0.90**
	(0.121)	—	(0.162)	(0.248)	(0.369)
Panel C: Extensive Broker FE (within-broker LPM, pp)					
$\hat{\beta}$	0.004	−0.004	0.004	0.006	0.006
	(0.001)	(0.000)	(0.002)	(0.002)	(0.005)
Effect (pp)	0.43***	−0.37	0.43***	0.59**	0.59
	(0.120)	(0.372)	(0.160)	(0.243)	(0.377)
Panel D: IV-Extensive (2SLS, pp)					
$\hat{\beta}$	0.014	−0.008	0.014	0.018	0.019
	(0.003)	(0.010)	(0.004)	(0.006)	(0.008)
Effect (pp)	1.41***	−0.76	1.42***	1.83***	1.88**
	(0.276)	(0.957)	(0.368)	(0.568)	(0.792)
Panel E: Int log (intensive log OLS, %)					
$\hat{\beta}$	0.115	−0.060	0.018	0.230	0.224
	(0.069)	(0.303)	(0.098)	(0.136)	(0.144)
Effect (%)	12.21*	−5.79	1.80	25.88*	25.06
	(7.75)	(28.58)	(9.93)	(17.14)	(18.06)
N customers	164,658	13,335	91,682	37,798	21,843
N brokers	1,063	132	594	217	119
Segment FE	Y	Y	Y	Y	Y
First-day FE	Y	Y	Y	Y	Y
Broker-rank FE	Y	–	–	–	–
Broker-cohort FE	Y	Y	Y	Y	Y

Table A9: Full regression output: Premium-15d, by broker seniority

Notes: Raw regression coefficients ($\hat{\beta}$) with broker-clustered standard errors in parentheses. Effects: % for PPML and Int log; pp for Extensive, Extensive Broker FE, and IV-Extensive. "—" indicates that cluster-robust standard errors do not converge in the L1 subsample (132 brokers). Int log subsample: N = 7,290 pooled, 431 (L1), 3,888 (L2), 1,723 (L3), 1,248 (L4+). Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

	Pooled	L1	L2	L3	L4+
Panel A: PPML (%, level effect)					
$\hat{\beta}$	0.264	0.374	0.283	0.221	0.277
	(0.092)	(0.677)	(0.129)	(0.182)	(0.200)
Effect (%)	30.23***	45.37	32.75**	24.79	31.88
	(12.01)	(98.43)	(17.15)	(22.77)	(26.34)
Panel B: Extensive (LPM, pp)					
$\hat{\beta}$	0.006	—	0.006	0.007	0.008
	(0.001)	—	(0.001)	(0.002)	(0.003)
Effect (pp)	0.58***	—	0.61***	0.71***	0.78**
	(0.103)	—	(0.137)	(0.203)	(0.336)
Panel C: Extensive Broker FE (within-broker LPM, pp)					
$\hat{\beta}$	0.004	−0.003	0.004	0.005	0.005
	(0.001)	(0.003)	(0.001)	(0.002)	(0.003)
Effect (pp)	0.41***	−0.26	0.44***	0.50**	0.49
	(0.102)	(0.310)	(0.135)	(0.196)	(0.344)
Panel D: IV-Extensive (2SLS, pp)					
$\hat{\beta}$	0.013	−0.006	0.014	0.016	0.016
	(0.002)	(0.008)	(0.003)	(0.005)	(0.007)
Effect (pp)	1.32***	−0.59	1.40***	1.56***	1.63**
	(0.237)	(0.804)	(0.315)	(0.490)	(0.681)
Panel E: Int log (intensive log OLS, %)					
$\hat{\beta}$	0.154	−0.289	0.096	0.275	0.207
	(0.082)	(0.320)	(0.115)	(0.172)	(0.183)
Effect (%)	16.68*	−25.10	10.09	31.66	22.96
	(9.63)	(23.94)	(12.67)	(22.63)	(22.52)
N customers	164,658	13,335	91,682	37,798	21,843
N brokers	1,063	132	594	217	119
Segment FE	Y	Y	Y	Y	Y
First-day FE	Y	Y	Y	Y	Y
Broker-rank FE	Y	–	–	–	–
Broker-cohort FE	Y	Y	Y	Y	Y

Table A10: Full regression output: Premium-1d, by broker seniority

Notes: Raw regression coefficients ($\hat{\beta}$) with broker-clustered standard errors in parentheses. Effects: % for PPML and Int log; pp for Extensive, Extensive Broker FE, and IV-Extensive. "—" indicates that cluster-robust standard errors do not converge in the L1 subsample (132 brokers). Int log subsample: N = 5,280 pooled, 300 (L1), 2,813 (L2), 1,263 (L3), 904 (L4+). Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

C.3 Dialogue-round outcomes, by customer segment and broker seniority

	Pooled	B	A	S	S+	New
Panel A: Dialogue rounds — PPML (%, level effect)						
$\hat{\beta}$	0.086	0.098	0.073	0.134	0.070	0.122
	(0.013)	(0.019)	(0.019)	(0.046)	(0.049)	(0.071)
Effect (%)	9.02***	10.31***	7.54***	14.32***	7.27	12.97*
	(1.39)	(2.09)	(2.08)	(5.30)	(5.21)	(7.97)
Panel B: Dialogue rounds — Int log (%)						
$\hat{\beta}$	0.059	0.052	0.063	0.082	0.045	0.065
	(0.006)	(0.009)	(0.009)	(0.028)	(0.028)	(0.038)
Effect (%)	6.07***	5.31***	6.53***	8.53***	4.60	6.68*
	(0.66)	(0.96)	(0.93)	(3.05)	(2.97)	(4.05)
Panel C: Pre-buy rounds — PPML (%, purchaser subsample)						
Effect (%)	−1.09	3.07	−11.62*	7.25	0.18	52.01***
	(4.59)	(8.60)	(6.73)	(12.14)	(11.19)	(23.24)
N customers	164,658	61,571	81,871	8,078	8,727	4,411
N brokers	1,063	1,049	1,057	1,004	1,005	452
Segment FE	Y	–	–	–	–	–
First-day FE	Y	Y	Y	Y	Y	Y
Broker-rank FE	Y	Y	Y	Y	Y	Y
Broker-cohort FE	Y	Y	Y	Y	Y	Y

Table A11: Full regression output: dialogue rounds, by customer segment

Notes: Panels A–B: raw regression coefficients ($\hat{\beta}$) with broker-clustered standard errors in parentheses; effects = $100(e^{\hat{\beta}} - 1)$. Panel C (Pre-buy rounds): only effects and standard errors are reported, as this outcome uses a single PPML specification on the purchaser subsample ($N = 5{,}280$ same-day buyers; $N_{\text{brokers}} = 975$). Pre-buy rounds is defined only for customers with positive same-day premium. Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

	Pooled	L1	L2	L3	L4+
Panel A: Dialogue rounds — PPML (%, level effect)					
$\hat{\beta}$	0.086	0.037	0.080	0.099	0.115
	(0.013)	(0.048)	(0.018)	(0.022)	(0.036)
Effect (%)	9.02***	3.80	8.33***	10.43***	12.20***
	(1.39)	(5.00)	(1.93)	(2.43)	(4.03)
Panel B: Dialogue rounds — Int log (%)					
$\hat{\beta}$	0.059	—	0.057	0.052	0.080
	(0.006)	—	(0.008)	(0.012)	(0.018)
Effect (%)	6.07***	—	5.87***	5.33***	8.37***
	(0.66)	—	(0.90)	(1.28)	(1.94)
Panel C: Pre-buy rounds — PPML (%, purchaser subsample)					
Effect (%)	−1.09	−29.93*	−3.37	11.39	4.95
	(4.59)	(15.90)	(6.50)	(8.85)	(10.70)
N customers	164,658	13,335	91,682	37,798	21,843
N brokers	1,063	132	594	217	119
Segment FE	Y	Y	Y	Y	Y
First-day FE	Y	Y	Y	Y	Y
Broker-rank FE	Y	–	–	–	–
Broker-cohort FE	Y	Y	Y	Y	Y

Table A12: Full regression output: dialogue rounds, by broker seniority

Notes: Panels A–B: raw regression coefficients ($\hat{\beta}$) with broker-clustered standard errors in parentheses; effects = $100(e^{\hat{\beta}} - 1)$. "—" indicates that cluster-robust standard errors do not converge in the L1 subsample (132 brokers). Panel C (Pre-buy rounds): only effects and standard errors are reported, as this outcome uses a single PPML specification on the purchaser subsample (N = 5,280 same-day buyers). Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.