Directional AI Advice: Experimental Evidence from Healthcare*

Yuyu Chen† Hongbin Li‡ Lingsheng Meng§ Xinyao Qiu¶ Qingxu Yang‖

September 7, 2026

Abstract

Generative AI is fast becoming the first place people turn for expert advice. The advice it provides can be directional rather than neutral, shaped in part by the choices of its designers and regulators. When clients consult AI before meeting an expert, they carry this directional advice into a relationship that once rested on the expert's judgment alone. We study its consequences in healthcare through a large-scale preregistered field experiment at a Chinese hospital, where we randomize patients' access to an AI chatbot before their outpatient visit. Examination of the conversation logs shows that the chatbot routinely cautions against the use of medications, especially Traditional Chinese Medicine and antibiotics, while issuing clean recommendations for diagnostic testing, consistent with the liability-driven guardrails encoded in AI training. This directionality propagates into clinical practice. Prescription rates decline among treated patients while diagnostic testing increases, and these effects are more pronounced among physicians who are receptive to patient input and those with more intensive prescribing styles. Beyond shifting healthcare utilization, survey results show that AI access reduces patient satisfaction and intended compliance, changing the decision-making relationship between patients and physicians.

Keywords: Generative AI, Credence Goods, AI Guardrails, Healthcare Field Experiment.

JEL Codes: I11, I18, O33, O38, D83.

*This experiment was preregistered at AEA RCT Registry (AEARCTR-0015851), and approved by Institutional Review Board of Guanghua School of Management, Peking University (#2025-14). We are grateful for helpful comments from Naijia Guo, Jin Li, Katrine Løken, Bentley MacLeod, Sultan Mehmood, Juan Pantano, Uta Schönberg, Yanhui Wu, Liang Zhong, Xiaodong Zhu, and seminar participants at the NBER Summer Institute 2026, SITE 2026, and the University of Hong Kong.

†Guanghua School of Management, Peking University. Email: chenyuyu@gsm.pku.edu.cn

‡Stanford University. Email: hongbinli@stanford.edu

§Stanford University. Email: lmeng@stanford.edu

¶Faculty of Business and Economics, The University of Hong Kong. Email: xinyaoq@hku.hk

‖School of Economics, Fudan University. Email: qxyang@fudan.edu.cn

1 Introduction

Generative artificial intelligence (AI) is becoming the first place many people turn before they turn to experts. Unlike traditional information technologies such as newspapers and search engines, generative AI goes beyond providing easier access to information. It also interprets, exercises judgment, and often recommends a course of action. These recommendations may not be neutral. AI's responses reflect the priorities of those who build and govern these systems.[1] For example, to limit legal exposure, developers install guardrails that steer their models away from liability-prone recommendations and toward more cautious alternatives. This directionality matters because AI now reaches into decisions people once delegated entirely to experts, in domains such as healthcare, law, and finance. As more people consult AI before turning to an expert, these design choices can propagate into real decisions at scale.

This paper examines whether and how directional advice from AI shifts the decisions experts and clients reach, in the context of healthcare. Healthcare is a canonical credence-good market characterized by severe informational asymmetries and substantial expert discretion (Arrow, 1963; Darby and Karni, 1973; McGuire, 2000), and it has now become a common use case for consumer-facing generative AI tools.[2] Its high stakes have drawn attention from both policymakers and developers, with developers implementing cautious guardrails to protect patients from the risks of acting on AI medical advice and to limit legal exposure. When patients consult AI before seeing a physician, the information they receive narrows the asymmetry that underlies physician discretion. But that information is not neutral. It also carries the priorities of AI developers, which may diverge from those of the health system.

To test how AI advice propagates into practice, we conduct a preregistered, month-long field experiment that provides the first randomized evidence on how patients' access to generative AI reshapes clinical decisions and the patient-physician relationship. Our randomized controlled trial takes place in a large public hospital in China and involves over 10,000 outpatient visits. Patients are randomly assigned to receive, or not receive, access to an LLM-based chatbot prior to their

[1] Pope Leo XIV's 2026 encyclical *Magnifica Humanitas* similarly emphasizes that artificial intelligence cannot be treated as a neutral force, because it reflects the aims of those who build, fund, regulate, and use it, and that its deployment reshapes relations of power and responsibility.

[2] Health-related queries are among the three most common use cases of ChatGPT (https://openai.com/index/introducing-gpt-5/).

appointments. The chatbot invites patients to describe their symptoms and health concerns in detail before the visit and provides free access to AI-generated, personalized guidance and general medical information. We implement a two-layer randomization design. We first randomize physicians into an *Exposed* group or an *Unexposed* group. Among patients scheduling appointments with *Exposed* physicians, half are randomly assigned to receive access to the chatbot (*Treatment*) and half to receive no access (*Control*). Patients scheduling appointments with *Unexposed* physicians never receive chatbot access. This design allows us to compare clinical outcomes between patients with and without access to the chatbot who see the same physician, while also testing for within-physician spillovers in clinical practice affecting patients without access.

We first examine take-up of the chatbot and the content of the advice it provides. Linking usage logs to hospital administrative records, we find that 17 percent of patients offered access used it before their visit, with higher take-up among younger, male, and employed patients and first-time visitors. Turning to the content of these interactions, we find a sharp asymmetry in the advice the chatbot provides. When discussing medications, the chatbot almost always attaches cautions, warning about side effects, contraindications, and the risks of self-medication. Such cautions appear in 69.7% of general medication mentions, 90.8% of Traditional Chinese Medicine (TCM) mentions, and 87.6% of antibiotic mentions, while clean recommendations are rare, at 3.8% for TCM and 7.8% for antibiotics. Diagnostic testing is treated very differently, with the chatbot issuing clean recommendations 94.5% of the time. This asymmetry tracks the liability concerns of LLM developers. Recommending a specific medication carries substantial legal risk if it leads to adverse health outcomes, whereas encouraging diagnostic testing does not, and the cost of unnecessary tests falls on patients and the health system rather than the developer. The advice patients carry into the consultation is therefore directional, steering away from medication and toward diagnostic testing.

This directionality propagates into clinical decisions.[3] Within-physician comparisons show that chatbot access reduces the likelihood of receiving any prescription by 4.6 percentage points relative to a baseline of 87%, driven partly by declines in TCM and antibiotic prescribing. At the same time,

[3]By directionality, we mean that the chatbot cautions against medications and encourages diagnostic testing, and that access moves clinical decisions the same way. Our randomization design identifies the direction of the effects on clinical decisions; the asymmetric advice behind it is descriptive, measured on the conversation logs of adopters, and consistent with defensive guardrails. Attributing the directionality to defensive guardrails specifically, rather than to other correlated features of model training, would require an experiment that randomizes patients across models with different guardrails.

chatbot access increases the probability that a visit includes diagnostic testing by 2.7 percentage points from a baseline of 23%. The direction of each treatment effect mirrors the direction of the AI's advice. These shifts lower medication expenditures and modestly raise diagnostic spending, leaving total healthcare expenditures unchanged. We find no evidence of worse short-term outcomes; if anything, treated patients are 1.2 percentage points less likely to revisit the same hospital within two weeks, though this estimate is only marginally significant and may reflect care-seeking behavior rather than improvements in health.

Our main results report intent-to-treat effects based on randomized access to the chatbot, regardless of actual usage. We additionally estimate treatment-on-the-treated effects by instrumenting actual chatbot usage with randomized assignment. Chatbot usage reduces the likelihood of receiving any prescription by 27 percentage points, increases diagnostic testing by 16 percentage points, and lowers two-week revisits by 7.0 percentage points. These effects are roughly five times larger than the corresponding intent-to-treat estimates, reflecting the partial take-up rate of 17%, and should be interpreted as local average treatment effects for patients who choose to use the chatbot when access is provided.

The effects are concentrated among the physicians for whom the intervention has the greatest scope to alter decisions. Because the chatbot operates by changing patients' information and the requests they make, its influence should depend on a physician's receptiveness to patient input and on the extent of baseline prescribing that AI-assisted patients might constrain. We find that the reduction in prescribing and the increase in testing are concentrated among physicians who self-report being open to patient input, and largely absent among physicians who emphasize strict authority. The reductions are also larger among physicians who prescribe most intensively at baseline, whether measured by their stated attitudes toward these medications or by their actual pre-experiment prescribing.

We further explore whether the effects spill over or persist, either to other patients or beyond the period of the experiment. One possibility is that physicians carry the effect. If repeated interactions with AI-prepared patients led physicians to adjust their practice style, we would expect within-physician spillovers to patients without chatbot access during the experiment, as well as persistence in physician practice after the experiment ends. We find neither. Among patients without chatbot access, clinical outcomes do not differ between those who see *Exposed* physicians and those who see

Unexposed physicians. We also find no persistence of effects: differences in clinical practice across physician groups disappear immediately after the experiment ends. This null result is economically informative. It suggests that generative AI does not automatically reshape physician practices, at least at the exposure rate in our experiment.

A second possibility is that patients carry the effect. While the intervention produces no lasting change in physician practice, its effects on patients may be more persistent. Following the same patients in the months after the experiment, we find that the increase in diagnostic testing remains, at 1.6 percentage points, while the reduction in prescribing fades. The evidence is suggestive, but it indicates that a single exposure to the chatbot may shift patient behavior in ways that outlast direct access. Taken together, these findings indicate that the chatbot's effects operate through patients' own use and do not diffuse to other patients through changes in physician practice. From a policy perspective, this implies that unequal access to AI tools may translate directly into unequal effects.

Finally, to understand how patient access to the chatbot affects patient behavior and the physician–patient relationship, we combine administrative medical records with survey data from patients and physicians. We find that access to the chatbot does not materially change observable patient behavior, including appointment attendance, medication purchases, or completion of diagnostic tests. However, survey evidence points to a divergence in perceptions of the consultation experience, despite several individual estimates being imprecise given the smaller survey samples. Patients with chatbot access report lower overall satisfaction, perceive communication with physicians as less smooth, and express lower intended compliance with medical advice. Physicians who interact with AI-prepared patients also report lower patient compliance, but at the same time perceive these patients as easier to communicate with, with clearer symptom descriptions and a better understanding of their condition. The reduction in both intended and observed compliance is consistent with Finkelstein et al. (2022) and suggests that generative AI may shift decision-making authority within clinical encounters, reshaping how patients view physician expertise and potentially weakening adherence.

The welfare implications are difficult to assess. The intervention reduces prescribing in categories where overuse is well-documented and increases diagnostic testing, a pattern consistent with welfare improvement if overtreatment is the relevant margin. However, we cannot rule out that some reduced prescriptions represent clinically appropriate care, or that the added testing generates costs without matching benefits. The reduction in compliance is similarly ambiguous. Lower adherence can spare

patients unnecessary treatment but can also lead them to decline beneficial care. We do not attempt to sign the net welfare effect. We emphasize instead that the direction of these changes reflects defensive guardrails encoded in AI training, and that the objectives AI developers optimize for need not align with those of the health system.

This study makes multiple contributions to the economics literature. First, and most centrally, we provide the first large-scale evidence on a question that has received little attention: how do the norms and constraints encoded in AI training propagate into real-world decisions at scale? LLMs exhibit systematic behavioral patterns shaped by their design, including tendencies that vary with model architecture, training data, and scale (e.g., Filippas et al., 2024; Ross et al., 2024; Bowen et al., 2025; Garg and Fetzer, 2025; Bini et al., 2025; Cheung et al., 2025; Conlon and Schwardmann, 2026; Waight et al., 2026). In high-stakes domains such as healthcare, these patterns are not incidental. Major AI developers have explicitly encoded defensive, liability-driven guardrails into their models, steering them away from specific treatment recommendations and toward caution and diagnostic clarification.[4] When such models are deployed as consumer-facing information tools, these encoded constraints enter real-life decisions at scale.[5] Our findings suggest this propagation mechanism is quantitatively important, with the potential to shift aggregate outcomes such as population-level healthcare utilization.

More broadly, we contribute to the rapidly growing economics of AI literature. Much of the existing work studies generative AI as a production technology that augments worker productivity or substitutes for labor in specific tasks (e.g., Dell'Acqua et al., 2023; Peng et al., 2023; Noy and Zhang, 2023; Cui et al., 2024; Brynjolfsson et al., 2025; Dillon et al., 2025; Reimers and Waldfogel, 2026). Viewed through the lens of expert–client relationships, a smaller stream examines AI tools that assist experts themselves, including physicians (Li et al., 2024; Abaluck et al., 2026; He and Xia, 2026),[6] financial analysts (Bertomeu et al., 2025), and legal workers (Choi et al., 2024; Mehmood et

[4] Chan (2026) provides a theoretical framework to study medical liability when artificial intelligence acts as a doctor.

[5] Luettgau et al. (2025) shows that a large majority of users follow personal advice from AI chatbots, even in high-stakes contexts.

[6] Recent randomized trials in clinical settings have tested integrating LLM tools into hospital workflows. Wan et al. (2024) find that an LLM chatbot assisting nurses at outpatient reception improves patient satisfaction and communication efficiency, and Tao et al. (2026) show that a pre-consultation chatbot performing clinical history-taking, preliminary diagnoses, and test ordering reduces specialist consultation time and improves coordination. Both studies deploy AI to augment provider functions and find improved patient satisfaction. In contrast, our study deploys AI as a patient-side information tool and examines its effects on both clinical decisions and the broader dynamics of the physician–patient relationship.

al., 2026). Much less is known about generative AI as a tool for consumers who face informational disadvantages relative to the experts they rely on. The limited existing work includes Ash et al. (2025), who randomize voters' access to AI-powered chatbots that answer questions about ballot initiatives, Shah and Levy (2026), who document that generative AI has substantially increased self-representation in U.S. federal civil courts, and Holz et al. (2026), who randomize household access to an AI chatbot for filing property tax appeals.[7]

Next, we contribute to a literature on the role of patient information in healthcare. Existing studies typically examine information asymmetries in healthcare by exploiting whether patients are physicians themselves (Grytten et al., 2011; Johnson and Rehavi, 2016; Frakes et al., 2021; Chen et al., 2025) or have family members in the medical profession (Artmann et al., 2022; Chen et al., 2022; Finkelstein et al., 2022) as sources of variation in patients' information, but the evidence from this literature is mixed. Some studies find that better-informed patients experience less overtreatment (Johnson and Rehavi, 2016), while others report only limited effects (Frakes et al., 2021; Artmann et al., 2022). By leveraging LLM-based technology, our study contributes to this body of literature by randomly providing a subset of patients with access to personalized, easy-to-understand, but potentially directional advice on their health conditions.[8] More broadly, the project also relates to studies of health-related information and education programs in developing country settings, as surveyed in Dupas (2011b).

Going beyond clinical decisions, we examine how access to information reshapes patient–physician interactions. Similar to Currie et al. (2011), who show that knowledgeable simulated patients constrain physicians' propensity to overprescribe antibiotics but experience less courteous interactions, we find that patients' access to LLM tools reduces prescribing while worsening self-reported experiences. At the same time, patients with access to LLM tools report lower compliance with physician recommendations, consistent with observational evidence that patients with family members in the medical profession exhibit reduced adherence (Finkelstein et al., 2022). These findings highlight

[7]There are also studies conducted in online experimental settings, including Spatharioti et al. (2023), which uses an online experiment to examine how LLM-based tools help consumers search for and compare products, and Choi et al. (2026) and Choukhmane et al. (2026), which examine AI-generated financial advice and its implications for household portfolios through online surveys and experiments.

[8]Related literature studies other types of health information, including second opinions (Mimra et al., 2016), patient decision aids that provide standardized materials about specific treatment options (surveyed by Stacey et al., 2017, 2024), provider report cards and online ratings that disclose information about healthcare provider quality (e.g., Dranove et al., 2003; Kolstad, 2013; Angerer et al., 2026), and educational campaigns on health risks and the importance of diagnostic testing (e.g., Dupas, 2011a; Lopez et al., 2024).

the importance of looking beyond treatment and testing decisions to understand how information technologies reshape relational dynamics in clinical care.

Finally, our study contributes to a broad literature on credence goods markets. These markets are economically significant and pervasive. From real estate (Levitt and Syverson, 2008) and financial advice (Mullainathan et al., 2012) to pharmacy (Bronnenberg et al., 2015), automobile and computer repair (Schneider, 2012; Busse et al., 2017; Rasch and Waibel, 2018; Kerschbamer et al., 2023), and taxi services (Balafoutas et al., 2013, 2017), consumers routinely delegate decisions to better-informed experts whose recommendations they cannot easily verify. A common thread across these settings is that informational asymmetries between experts and clients create scope for overtreatment, fraud, or misallocation. Consistent with this, Balafoutas et al. (2013), Busse et al. (2017), and Kerschbamer et al. (2023) show that reducing informational asymmetries can discipline expert behavior. Our study extends this literature by showing that generative AI can serve as a scalable consumer-side information technology, shifting expert–client relationships in one of the most consequential credence goods markets: outpatient healthcare.

2 Institutional Background

2.1 Generative AI in Healthcare: Adoption and Guardrails

Generative AI has rapidly become a major channel through which people access health information. Health-related queries are among the most common use cases of large language models: OpenAI reports that approximately 230 million people discuss their health with ChatGPT each week,[9] and surveys indicate that roughly one in six American adults consults an AI chatbot for health information at least once a month (Kaiser Family Foundation, 2024). This growth has occurred largely ahead of formal regulatory frameworks, creating a setting in which AI systems interact with patients at massive scale without the oversight structures that govern licensed medical professionals.[10]

In response to safety concerns and legal exposure, major AI developers have encoded defensive

[9] https://openai.com/index/introducing-chatgpt-health/.

[10] The regulatory debate over AI's role in healthcare is actively evolving. California's Assembly Bill 489, signed into law in October 2025, directly addresses the boundary between AI and licensed medical practice: it prohibits developers and deployers of AI systems from using any terms, phrases, or design elements that imply the system is providing care from a licensed healthcare professional, and authorizes state licensing boards to investigate and penalize violations. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260AB489.

guardrails into their models for health-related interactions. The liability motive is stated explicitly. OpenAI's published guidelines list protecting the company from legal and reputational harm as one of the objectives of model behavior.[11] On medical matters, the guidelines instruct the models to provide information but not definitive advice, to include disclaimers directing users to licensed professionals, and, for mental health, to refrain from providing diagnoses or recommending specific treatments or medications. Anthropic has similarly stated that Claude is designed to include contextual disclaimers, acknowledge uncertainty, and direct users to healthcare professionals for personalized guidance.[12] Audit studies confirm that these stated policies manifest in model behavior: disclaimers and referral advice scale with the clinical urgency of the query, differ across developers, and shift in form over time (e.g., Reis et al., 2026; Sharma et al., 2025).

Because general-purpose AI models are not subject to the pre-market review or professional oversight that governs medical practice, these guardrails function as self-imposed substitutes for formal regulation, and their content reflects the objectives of developers, which need not coincide with those of the health system. The restraint is also uneven. The guardrails restrict the actions most likely to create liability for the developer, recommending treatments or medications and providing diagnoses, while directing users toward licensed professionals and imposing no comparable restraint on encouraging further evaluation. When patients consult these tools, this defensive and potentially directional logic reaches them at scale, and may reshape their healthcare decisions.

2.2 Healthcare Market and Institutional Context in China

Healthcare expenditures have risen rapidly worldwide. Economic research points to both demand-side and supply-side mechanisms as drivers of high healthcare spending. On the demand side, beyond the classic moral hazard problem,[13] patients often lack the medical knowledge needed to evaluate the necessity of care and may request treatments that are not clinically indicated. Physicians may comply with such requests, contributing to overutilization (e.g., Kravitz et al., 2005; Kotwani et al., 2010; Linder et al., 2014; Finkelstein et al., 2016; Lopez et al., 2022). On the supply side, physician

[11] https://model-spec.openai.com/2025-04-11.html. We cite the version in effect during our experimental period.

[12] https://www.anthropic.com/news/healthcare-life-sciences.

[13] China has near-universal health insurance coverage through two main public schemes: the Urban Employee Basic Medical Insurance (UEBMI), which covers formal-sector workers, and the Urban and Rural Resident Basic Medical Insurance (URRBMI), which covers rural residents and non-working urban populations.

characteristics and financial incentives directly shape treatment decisions, generating systematic variation in medical expenditures, including overspending and inefficiencies (e.g., Iizuka, 2012; Das et al., 2016; Finkelstein et al., 2016; Cutler et al., 2019; Badinski et al., 2023; Currie et al., 2026).[14]

Overutilization, particularly in the form of overprescription, is well documented in China (e.g., Li et al., 2012; Lu, 2014; Fang et al., 2021). The healthcare system in China is predominantly state-run, with public hospitals delivering the vast majority of services. Despite ongoing payment reforms, hospitals continue to operate largely under a fee-for-service payment model and receive limited direct government subsidies. As a result, hospital revenue remains closely tied to service volume. Pharmaceutical sales occur primarily within hospitals rather than through independent retail pharmacies. Between 2014 and 2019, drug expenditures represented more than 60 percent of inpatient spending and roughly 70 percent of outpatient spending in public hospitals.[15] These shares are substantially higher than those observed in many high-income health systems, where prescription drugs account for a considerably smaller fraction of total health expenditures.[16]

Within this broad pattern of overprescription, three drug categories draw particular policy scrutiny: Traditional Chinese Medicine (TCM), antibiotics, and opioids. TCM is widely prescribed in Chinese public hospitals but is characterized by less standardized evidence on effectiveness and side effects,[17] and often higher or more variable pricing. Antibiotic overprescription has been well documented in China (Currie et al., 2013, 2014) and in the United States (Fleming-Dutra et al., 2016), raising concerns about antimicrobial resistance. Opioids, while less prevalent in China than in the United States (Maclean et al., 2020), have drawn substantial international attention due to concerns about misuse and addiction. We focus on these three categories because they are highly policy relevant and have been central to regulatory efforts and public health debates.

Motivated by this institutional setting, we examine how patient access to generative AI affects clinical decisions in China. With access to AI, patients can arrive better informed, less likely to request unnecessary treatments, and better positioned to constrain physicians' overprescription. The

[14] Moreover, Gottschalk et al. (2020) document that patients with higher socioeconomic status are significantly less likely to receive unnecessary treatment recommendations, indicating that physicians' propensity to overtreat depends on patient characteristics.

[15] China Health Statistical Yearbook, various years.

[16] Across OECD countries, pharmaceutical spending typically accounts for roughly 15–20% of total health expenditures. https://www.oecd.org/en/data/indicators/pharmaceutical-spending.html.

[17] Under new regulations effective July 1, 2026, TCMs that do not clearly disclose contraindications, adverse reactions, or precautions will not be approved upon re-registration. Reports indicate that over 70 percent of currently approved products may be affected. https://www.news.cn/politics/20260129/853dc80d18fb43d2b1d61676c46b5675/c.html.

information they obtain may not be neutral, however. Shaped by developer guardrails, it may push clinical decisions away from medications and toward diagnostic testing rather than simply reducing unnecessary care uniformly.

Patient access to AI is likely to be especially consequential in the Chinese outpatient setting, where patients have few institutional sources of guidance. In contrast to systems organized around general practitioners, patients in China schedule directly with specialists, without referral or triage by a family doctor who helps interpret their condition or provides continuity of guidance across visits.[18] Outpatient consultations are brief, averaging roughly nine minutes per visit in our partnering hospital, which constrains explanation, counseling, and patient engagement. With neither an intermediating generalist nor sufficient consultation time, patients have little means of obtaining or interpreting medical information on their own. In this setting, the potential demand for AI advice is high, and so is the scope for its influence on the consultation.

3 Experimental Design and Data

We conducted the experiment in partnership with a large public hospital in Sichuan Province. The hospital employs over 200 outpatient physicians who handle approximately 900 visits per day on average. As a city-level tertiary hospital, it is representative of the mid-tier public institutions that serve roughly 40% of China's population and account for 35% of nationwide outpatient visits, supporting the external validity of our findings. The patient population is predominantly rural and has limited exposure to digital technologies, providing a low baseline of AI adoption that minimizes the risk of background usage diluting the treatment effect. The hospital's electronic medical record system captures detailed visit-level data on diagnoses, prescriptions, diagnostic test orders, and other clinical details. Its online appointment platform allows us to randomize chatbot access at the time patients book their appointments.

[18] In credence-good markets, repeated interaction is one of the few mechanisms that can discipline expert behavior even absent formal oversight (Dulleck and Kerschbamer, 2006; Fong et al., 2022). In the Chinese outpatient system, specialist visits offer no substitute for this missing continuity, as the same patient rarely sees the same specialist repeatedly.

3.1 Intervention

We partnered with a startup that specializes in AI solutions for the healthcare industry. Our intervention is a generative AI chatbot powered by a large language model built on a widely used general-purpose architecture, adapted for healthcare-related interactions, and integrated into the hospital's online appointment system. Immediately after a patient completes an online appointment booking, which occurs on average one day before the visit, those randomly assigned to the treatment group receive a personalized link to the chatbot. Simultaneously, the hospital sends a separate message encouraging patients to use the tool prior to their visit. Panels (a) and (b) of Figure 1 present the chatbot's initial landing screen and consultation interface.

The chatbot supports pre-visit patient preparation by eliciting symptom descriptions, interpreting and analyzing patient health concerns, and providing general medical information. Throughout all interactions, the interface prominently reminds users that the chatbot does not replace professional medical advice and that clinical decisions should follow physicians' recommendations. The system automatically logs all user interactions, including timestamps and message content. These records provide an objective measure of engagement with the intervention, which we link to administrative and clinical outcomes using unique patient identifiers. This structure enables estimation of both intent-to-treat and treatment-on-the-treated effects.

The intervention may affect both communication and clinical decision-making through several channels. First, by prompting patients to organize and describe symptoms and to review relevant health information before the visit, the chatbot can improve pre-visit preparation and facilitate clearer communication during the consultation. This is particularly important in our setting, where the average outpatient visit lasts approximately nine minutes, making efficient and precise communication essential. Beyond communication, access to structured medical information can improve patients' understanding of their health conditions, typical diagnostic pathways, and the role of medications and tests. Better-informed patients may therefore be less likely to request unnecessary treatments or prescriptions, reducing demand-driven overuse. In addition, the advice patients receive may not be neutral. Because guardrails encoded in AI training caution against medication use, patients may arrive carrying advice that tilts away from medications and toward diagnostic testing. Finally, by narrowing informational gaps, the chatbot may shift the decision-making authority within the

consultation. Patients who arrive with greater knowledge and more structured questions are better positioned to question recommendations and constrain physicians' discretion to overprescribe or overtreat. Through these channels, the intervention can reshape the clinical interaction between patients and physicians.

3.2 Two-Layer Randomization

We implement a two-layer randomized experimental design to identify the direct effects of providing patients with access to the chatbot while creating variation to detect within-physician spillovers. We first randomize physicians into an *Exposed* group (two-thirds) or an *Unexposed* group (one-third), stratified by seniority, education, and specialty. Among patients scheduling appointments with *Exposed* physicians, half are randomly assigned to the *Treatment* group, with chatbot access, and half to the *Control* group, without access to the chatbot. Patients scheduling appointments with *Unexposed* physicians do not receive chatbot access, so *Unexposed* physicians never interact with AI-assisted patients.

Panel (c) of Figure 1 summarizes the two-layer randomization. This design generates three analytically distinct groups. Comparisons between the *Treatment* and *Control* groups identify the direct effect of patient access to the chatbot holding the physician fixed. Comparisons between the *Control* and *Unexposed* groups detect within-physician spillovers arising from physician learning or behavioral adjustment following repeated interactions with AI-prepared patients, since neither group has chatbot access and the two differ only in their physicians' exposure.

The three analytical groups are formally labeled as:

1. **Treatment:** Patients with chatbot access, seeing *Exposed* physicians.
2. **Control:** Patients without chatbot access, seeing *Exposed* physicians.
3. **Unexposed:** Patients without chatbot access, seeing *Unexposed* physicians.

3.3 Data and Sample

The experiment was conducted between June 2 and July 3, 2025, at our partnering hospital. Because the intervention is delivered through the online appointment system, we restrict the sample to patients who scheduled outpatient visits online during this period. Patients who walked in without

an online appointment, approximately one third of total outpatient visits, are excluded from both the experiment and the analysis sample.[19]

During the experimental period, 179 physicians provided outpatient care. Of these, 117 (two thirds) were randomized to the *Exposed* group and 62 (one third) to the *Unexposed* group. Panel A of Table 1 reports balance tests at the physician level and shows that the two groups are well balanced across observed characteristics, including age, gender, seniority, education, and specialty, with no statistically significant differences.

During the experimental period, 12,498 online appointments with *Exposed* physicians were randomized with equal probability into the *Treatment* or *Control* group at the time of booking, and 6,435 appointments were scheduled with *Unexposed* physicians.[20] Patients did not show up for 6.6 percent of these appointments. An unattended appointment generates no clinical record and therefore cannot enter the analysis. The analysis sample accordingly comprises 11,666 completed visits with *Exposed* physicians and 6,010 with *Unexposed* physicians. The unit of observation is the outpatient visit rather than the patient, so patients who visited multiple times during the experimental period contribute multiple observations. We show in Section 6 that attrition is balanced across groups, indicating that chatbot access did not affect the decision to attend, so the restriction to completed visits does not introduce differential selection.

Panel B of Table 1 reports balance tests at the visit level, comparing the *Treatment* and *Control* groups as well as the *Treatment* and *Unexposed* groups. Across these comparisons, we find no statistically significant differences in age, gender, occupation, first-time visit status, or advance booking time.[21] The average patient is approximately 45 years old, and a large share report working as farmers (44%) or being unemployed or retired (16%). This demographic profile suggests that baseline familiarity with generative AI tools is likely to be limited in this population. Appointments are booked, on average, approximately 25 hours in advance, giving treated patients adequate opportunity to use the chatbot if they choose to do so.

[19]We additionally exclude pediatric patients from the experiment. Approximately 0.6% of patients who booked an appointment but did not complete payment are also excluded from the analysis sample.

[20]Within the hospital's system, when a physician orders diagnostic tests, the subsequent review of results, any follow-up discussion with the physician, and any prescribing based on those results fall under the original registration, even when the results are unavailable on the day of the initial consultation and the patient comes back later. The full sequence is recorded as one visit rather than several.

[21]First-visit status equals one if the visit is the patient's first visit to the hospital during the experimental period.

Medical Records. We observe administrative outpatient medical records for all visits during the study period. These records include patient and physician identifiers, visit timestamps, diagnoses, detailed drug prescriptions, and orders for diagnostic examinations. For each prescribed drug and diagnostic test, we observe the listed price, which captures medical expenditures prior to insurance reimbursement. These data allow us to construct visit-level measures of clinical decision-making, treatment intensity, and spending outcomes.

Patient Survey. We invited patients to complete a brief post-consultation survey on the day of the visit or the following day, offering a monetary incentive of 2 to 10 RMB. The survey elicits patients' overall satisfaction with the visit, perceived quality of communication, and intended adherence to medical advice. For patients in the *Treatment* group, the survey includes additional questions on chatbot usage and perceived usefulness. In total, we collected 2,247 patient responses, corresponding to a 13 percent response rate. The sample includes 640 respondents from the *Treatment* group, 789 from the *Control* group, and 818 from the *Unexposed* group. Response rates are somewhat lower in the *Treatment* group than in the two comparison groups; we therefore interpret survey-based outcomes with caution.

Physician Survey. Following the conclusion of the one-month experiment, we surveyed all physicians who conducted outpatient consultations during the study period. Of 179 physicians, 171 completed the questionnaire, yielding a response rate of 95.5 percent. The survey elicits physicians' assessments of their patients along four dimensions: clarity of symptom descriptions, understanding of their own health conditions, ease of communication during the visit, and potential adherence to medical advice. Both *Exposed* and *Unexposed* physicians were also asked to estimate how frequently their patients used AI tools prior to visits.

Taken together, the physician and patient surveys complement the administrative data by validating treatment exposure and providing insight into how AI-provided information affects consultation dynamics and the patient experience.

3.4 Chatbot Take-up

Although all patients in the *Treatment* group received free access to the chatbot prior to their consultation, actual usage was voluntary. Linking the usage logs to the hospital administrative system, we find that the chatbot was used before 998 of the 5,828 visits in the *Treatment* group (17.1 percent). Patients in the *Control* group were not given access for their visit and therefore could not use the tool.[22] Appendix Table A2 reports first-stage estimates of the relationship between treatment assignment and chatbot usage. Assignment to the *Treatment* group increases the probability of chatbot use by 17 percentage points.

Take-up among treated patients is self-selected and varies systematically with observable characteristics. Appendix Table A3 shows that, conditional on being assigned access, younger patients, men, employed patients, and first-time visitors are significantly more likely to use the chatbot. This selective adoption pattern has two implications for the analyses that follow. First, because only a subset of treated patients actively use the chatbot, intent-to-treat estimates reflect the effect of offering access at this take-up rate. We also use random assignment as an instrument for actual usage to estimate treatment-on-the-treated effects in Section 4.4. Second, the demographic profile of adopters suggests that the benefits of generative AI tools may not be evenly distributed across patient populations, a distributional concern we revisit in Section 5.

4 Main Results

4.1 The Directional Content of Patient–AI Conversations

We begin by examining how adopters in the *Treatment* group used the chatbot and what their conversations reveal about the advice they received.

Our analytic sample consists of 1,192 conversation turns from 956 visits, where each turn comprises one patient message and the chatbot's reply, and a single conversation may span multiple turns.[23] Two-thirds of patient messages (66.7%) describe symptoms and request a diagnosis or underlying cause. The remainder inquire about treatment options (11.7%), logistical matters (6.6%),

[22] Because randomization occurs at the appointment level, a patient with multiple visits can be assigned to different groups across visits; chatbot links, however, are appointment-specific and expire after the visit.

[23] The usage logs record chatbot use before 998 visits. Conversations for 42 of these visits could not be matched to hospital records and are excluded, leaving 1,192 conversation turns from 956 visits.

lifestyle and self-care (5.5%), whether to undergo specific tests (4.9%), the interpretation of test results or prior diagnoses (3.1%), and a small residual of greetings and uninformative messages (1.4%). Appendix B.1 reports the full distribution of intent labels along with example messages and details the classification procedure.

We next examine how the AI engages with four clinically relevant topics that map directly onto our outcomes of interest: diagnostic testing, general medication (excluding TCM and antibiotics), TCM, and antibiotics. We identify which topics a response covers through sentence-level keyword matching, and a single response typically spans several of them. For each topic a response covers, we use GPT-4o to classify the AI's stance along two independent dimensions. The first is whether the response *recommends* the test or treatment, that is, whether it endorses or suggests its use. The second is whether the response raises a *caution*, such as a warning about risks or side effects, a contraindication, or advice against use. We keep the two dimensions separate because they sometimes co-occur. Beyond outright advising against a treatment or recommending a treatment or test with no reservation, the chatbot frequently recommends a treatment while warning about its side effects and cautioning against self-use. Separating recommendation from caution lets us distinguish a clean recommendation from a hedged one. Every mention therefore falls into one of four categories: recommend without caution, recommend with caution, caution without recommending, or neither. Appendices B.2 and B.3 provide the prompts and the full classification procedure, and Table 2 reports the results.

Two patterns stand out. First, it is the AI, not the patient, that typically introduces these topics into the conversation. Patients mention general medication in only 11.4% of turns and diagnostic testing in 8.5% of turns, yet the AI raises medication in 58.9% of replies and testing in 58.2% of replies. The chatbot is not merely responding to patient queries; it actively shapes the informational content that patients carry into the clinical encounter. Notably, the AI raises medication and testing at nearly identical rates, and mentions TCM and antibiotics far less often (11.0% and 18.2% of replies).

Second, conditional on raising a topic, the AI's stance differs sharply between medications and diagnostic testing. When the AI discusses medication, it usually raises cautions, warning about side effects, contraindications, dosing concerns, or the risks of self-medication. Such warnings appear in 69.7% of general medication mentions, 90.8% of TCM mentions, and 87.6% of antibiotic

mentions. Clean recommendations, those without any accompanying caution, are rare for medications, particularly for TCM (3.8%) and antibiotics (7.8%). Diagnostic testing is treated very differently: when the AI discusses testing, it issues clean recommendations in 94.5% of cases, with cautions appearing in only 3.3% of mentions.

These patterns are consistent with the incentives facing LLM developers in high-stakes domains. Safety guardrails in health-related interactions serve both to protect users from the risks of acting on AI medical advice without professional oversight and to limit developers' legal exposure. Both rationales weigh against recommending specific treatments or medications, where adverse outcomes carry clinical and legal consequences. Neither weighs against recommending diagnostic testing, whose costs fall on patients and the health system rather than on the developer. As a result, AI responses are highly cautious about medications and tend to encourage diagnostic testing, a stance that need not align with the objectives of the health system.

The directional pattern documented above could in principle be specific to the model our partner deployed rather than a general feature of LLM-based medical advice. To probe this, we submit the same first-turn patient messages to six widely used large language models: GPT-4o, GPT-5.5, DeepSeek-V3, Claude Sonnet 4.6, Qwen3.5-plus, and Gemini 3.5 Flash. To approximate our setting, we use two system prompts, one framing the model as a pre-visit assistant and one as a direct medical consultant, and classify the replies with the same topic-detection and stance-classification pipeline used for our chatbot.[24] Figures A1 and A2 report the share of clean recommendations by topic, along with the number of replies in which each topic is mentioned.

The comparison shows the same directional asymmetry across models. Every model, under both prompts, recommends diagnostic testing far more cleanly than medications, with testing recommended without caution in 92% to 99% of mentions, close to the rate for our chatbot. The degree of caution toward medications nonetheless varies substantially: clean-recommendation rates for medications range from 10% to 56% across models, differing across developers and even across models from the same developer.[25] Among the comparison models, our chatbot's stance is closest to

[24] Appendix B.4 details the procedure and prompts. Because we cannot reconstruct the follow-up turns of each conversation, the comparison uses first-turn messages only, and restricts our chatbot to its first-turn responses to match. The resulting baseline shares therefore differ slightly from Table 2, which uses the full analytic sample.

[25] GPT-4o, for example, recommends medications without caution more than half the time. This is partly driven by its frequent use of deferral phrasing such as "the doctor may prescribe…," which we classify as a clean recommendation even though it defers the decision to the physician.

that of GPT-5.5, a leading frontier model by OpenAI, across all four topics.

The variation across models is itself informative. If the stance toward medications were a neutral reflection of clinical evidence, it would not differ so widely across models built by different developers. That it does indicates that the strength of the caution reflects design choices rather than the underlying medical evidence. The information patients receive from the chatbot is, in this sense, not a neutral distillation of medical knowledge but a view shaped by the priorities of those who build and govern these systems. Treated patients carry the chatbot's advice into their visits. Whether the encoded caution toward medications and the encouragement of diagnostic testing propagate into clinical decisions is the question we turn to next, using the medical records.

An alternative explanation for the asymmetry is informational rather than directional. A short text message from patients may not provide enough information for our AI model to assess the condition, in which case recommending diagnostic clarification is the appropriate response to uncertainty rather than a reflection of encoded safety guardrails. Under this explanation, patients who describe their symptoms in more detail should receive fewer recommendations for testing and more substantive guidance on medication. Appendix Table A1 tests this by splitting patients' messages into terciles of message length. Moving from the shortest to the longest tercile, the chatbot discusses diagnostic testing more often rather than less, and recommends testing without caution at a higher rate. Cautions attached to medication do not soften with message length either. Although descriptive, the analysis provides suggestive evidence against the explanation that the asymmetry in AI stance toward testing and medication reflects the AI's lack of information.

4.2 Clinical Outcomes Across Experimental Groups

We now turn to the clinical records to test whether this directional advice is reflected in what happens at the visit. We graphically present clinical outcomes for the *Treatment*, *Control*, and *Unexposed* groups over the experimental period, beginning with prescribing behavior. Panel (a) of Figure 2 shows that access to the chatbot reduces the likelihood that a visit results in any prescription. The share of visits with at least one prescribed medication is 4.5 percentage points lower in the *Treatment* group than in the *Control* group (significant at the 5% level), where the baseline prescribing rate is 88.8%. The prescribing rate in the *Treatment* group is also significantly lower than in the *Unexposed* group, while rates in the *Control* and *Unexposed* groups are statistically

indistinguishable.

We next examine prescribing patterns across major medication categories, including Traditional Chinese Medicine (TCM), antibiotics, and opioids. Panels (b) and (c) of Figure 2 show sizable reductions in TCM and antibiotic prescribing in the *Treatment* group, with statistically significant differences relative to both comparison groups. Opioid prescribing is also lower among treated patients, but the differences relative to the *Control* and *Unexposed* groups are not statistically significant. This is expected given the low baseline rate of opioid prescribing in China, below 3% in both the *Control* and *Unexposed* groups, as shown in Panel (d).

In contrast, physicians seeing treated patients are more likely to order diagnostic tests. Panel (e) shows that 21–22% of visits in the *Control* and *Unexposed* groups involve at least one diagnostic test, compared with 24.7% in the *Treatment* group. The increase in testing is statistically significant in both comparisons. This pattern of reduced prescribing and increased diagnostic testing is consistent with the directionality in the AI advice we observe from the conversation logs.

We next assess whether these changes in clinical management are associated with differences in patient outcomes. One outcome observed in our administrative data is whether a patient revisits the same hospital within two weeks of the initial consultation. Panel (f) of Figure 2 shows that patients in the *Treatment* group are 1.1 percentage points less likely to revisit within two weeks than patients in the *Control* group, a 7% reduction that is statistically significant. The same conclusion holds when comparing treated patients to the *Unexposed* group, whose two-week revisit rate is similar to that of the *Control* group. The reduction in short-term revisits is consistent with improvements in the effectiveness of initial consultations, but it may also reflect other mechanisms. Better-prepared patients may feel less need to seek follow-up care even absent changes in clinical quality. We therefore do not interpret this reduction as direct evidence of improved clinical outcomes. In this rural setting, where patients often face substantial travel costs, reductions in revisits nonetheless have welfare implications, even if the underlying channels remain unclear.

4.3 Within-Physician Intent-to-Treat Effects

Because outcomes are similar in the *Control* and *Unexposed* groups, our main regression analyses restrict the sample to visits with *Exposed* physicians.[26] Within this sample, we estimate the

[26]We return to the *Unexposed* group in later analyses to formally test for spillovers.

causal effects of offering access to the LLM-based chatbot on clinical outcomes using the following specification:

$$Y_i = \alpha + \beta \, \text{Treated}_i + \gamma X_i + \delta_{p(i)} + \varepsilon_i \tag{1}$$

where i indexes an outpatient visit and $p(i)$ denotes the physician seen at that visit. A patient who books more than one appointment during the experiment period contributes more than one visit, and can fall into different treatment groups.[27] The outcome variable Y_i represents outcomes from the consultation: prescriptions, diagnostic tests, and healthcare expenditures. The key explanatory variable, Treated_i, is a binary indicator equal to one if visit i was randomly assigned to chatbot access prior to the consultation. We include a vector of patient characteristics X_i (age, gender, occupation, and first-time visit status) and physician fixed effects $\delta_{p(i)}$. The error term ε_i captures unobserved factors affecting individual visits. All regressions are estimated using robust standard errors.[28]

The coefficient of interest, β, identifies the *Intent-to-Treat* (ITT) effect of being assigned chatbot access, regardless of actual usage. This approach preserves the integrity of the randomization and provides policy-relevant estimates of the effect of *offering* AI tools in a real-world clinical setting. It is important to note that if within-physician spillovers occur, for example, if physicians adjust their general practice patterns for all patients after interacting with AI-informed ones, the control group may also be indirectly affected. In that case, our estimate of β understates the true effect of chatbot exposure.

Table 3 reports within-physician intent-to-treat estimates of the effects of offering chatbot access on clinical practice patterns among visits with *Exposed* physicians. Consistent with the graphical evidence, the regression results show that access to the chatbot significantly reduces the likelihood that a visit results in any prescription by 4.6 percentage points relative to a sample mean of 87%. This

[27] Some patients visit the hospital more than once during the experiment period. To address the concern that the same patient could be assigned to different treatment conditions across visits, Appendix Table A6 restricts the sample to each patient's first visit during the experimental period, which retains 70% of visits. The main results are robust: chatbot access continues to significantly reduce overall prescribing and TCM prescribing and to increase diagnostic testing. The point estimates shift in line with the lower baseline prescribing rate and higher baseline testing rate of first visits, and the smaller estimates for antibiotics and two-week revisits are not statistically significant in this reduced sample.

[28] Standard errors are not clustered because randomization occurs at the individual appointment level (Abadie et al., 2023). Appendix Tables A7–A10 report estimates with standard errors clustered at the physician level. The core results on prescribing and diagnostic testing remain statistically significant. The two-week revisit effect is no longer marginally significant under clustering, and the reduction in Western Medicine expenditure is no longer significant once patient characteristics and physician fixed effects are added.

decline is driven partly by reductions in prescribing of Traditional Chinese Medicine and antibiotics, which fall by 2.2 and 1.5 percentage points, respectively, while opioid prescribing declines but the effect is not statistically significant. At the same time, chatbot access increases the probability that a visit involves any diagnostic test by 2.7 percentage points relative to a sample mean of 23%. Finally, treated patients are about 1.2 percentage points less likely to revisit the same hospital within two weeks; the effect is marginally statistically significant. Estimates are nearly identical with and without controls for patient characteristics and physician fixed effects, confirming that the randomization successfully balanced patient and physician characteristics.

In addition to whether a visit results in any prescription or any diagnostic test, Appendix Table A4 reports intent-to-treat estimates for the number of unique medications and diagnostic tests per visit. In the specification with patient controls and physician fixed effects, chatbot access reduces the number of prescribed medications by 0.160 relative to a sample mean of 2.44, a 6.6% decline, with reductions of 0.054 in TCM drugs (23% of the sample mean of 0.23), 0.015 in antibiotics (11% of 0.14), and 0.004 in opioids (29% of 0.014). The number of diagnostic tests ordered rises by 0.069 relative to a mean of 0.58, a 12% increase. The opioid coefficient is statistically significant at the intensive margin, while the corresponding extensive-margin estimate in Table 3 is not, consistent with the low baseline rate of opioid prescribing limiting power at the extensive margin.

Table 4 reports intent-to-treat effects on healthcare expenditures, measured as the sum of listed prices for prescribed drugs and diagnostic tests before insurance reimbursement. Consistent with the prescribing changes documented in Table 3, chatbot access reduces medication expenditures, with reductions in both TCM and Western Medicine expenditures significant at the 10% level. In the specification that controls for patient characteristics and physician fixed effects, TCM expenditures decline by approximately 1.76 RMB, a 13% reduction relative to the sample mean, while Western Medicine expenditures decline by about 5.0 RMB relative to a sample mean of 64.8 RMB. Diagnostic expenditures increase by roughly 6.3 RMB, though this estimate is not statistically significant. Because these changes offset, we find no statistically significant effect of chatbot access on total healthcare expenditures.

4.4 Treatment-on-the-Treated Effects

The results above report intent-to-treat effects based on randomized access to the chatbot, regardless of whether patients actively used the tool. We now estimate treatment-on-the-treated effects, which recover the average effect of chatbot usage among patients who complied with the assignment. We instrument actual chatbot usage with randomized assignment to the *Treatment* group, using the same within-physician specification as in the intent-to-treat analysis.

Table 5 reports the effects of chatbot usage on clinical practice patterns. Consistent with the intent-to-treat results, chatbot usage is associated with a 27 percentage point reduction in the probability that a visit results in any prescription, partly driven by declines of about 13 percentage points in Traditional Chinese Medicine prescribing and 9 percentage points in antibiotic prescribing. At the same time, chatbot usage increases the likelihood of diagnostic testing by 16.1 percentage points and reduces the probability of a two-week revisit by 7 percentage points. These magnitudes are approximately five times larger than the corresponding intent-to-treat estimates, reflecting partial take-up of the chatbot (17.1% in the *Treatment* group). The instrumental variable estimates therefore represent local average treatment effects for patients who use the chatbot when offered access. Overall, the scale of these effects suggests that generative AI tools can meaningfully alter clinical decision-making in ways that, if adopted broadly, could translate into substantial changes in healthcare utilization at the system level.

Table 6 reports the local average treatment effects on healthcare expenditures. Chatbot usage is associated with sizable reductions in medication expenditures, with TCM expenditures declining by 10 RMB and Western Medicine expenditures declining by roughly 30 RMB, both statistically significant at the 10% level. These reductions are offset by an increase in diagnostic expenditures of approximately 37 RMB, although this estimate is imprecisely estimated. For both Tables 5 and 6, first-stage F-statistics exceed 1,100 across specifications, indicating a strong first stage.

Chatbot usage is a coarse measure of exposure, since what patients take from the chatbot depends on the AI's responses to their questions. Appendix Table A5 disaggregates treated patients by the content of the advice they received, classifying each into one of four mutually exclusive categories *per topic* according to whether the chatbot provided only a recommendation, only a caution, both, or no

advice at all, with the *Control* group as the omitted category.[29] The direction of advice lines up with the direction of the clinical decision. Patients who received cautions without recommendations are substantially less likely to be prescribed the corresponding medication, by 21.0 percentage points for general medication, 18.4 for TCM, and 13.3 for antibiotics, while patients who received clean testing recommendations are 6.6 percentage points more likely to undergo diagnostic testing. The analysis is descriptive. Chatbot take-up is self-selected, the symptoms a patient describes determine the chatbot's responses, and those same symptoms shape the physician's decision, so a patient whose health condition prompts a clean medication recommendation may also be one the physician would have prescribed for. Cells corresponding to clean recommendations for TCM and antibiotics, and to cautions against testing, are also small. We therefore read the table as evidence on the correspondence between advice content and clinical decisions rather than as a causal estimate of the effect of advice content.

4.5 Heterogeneous Treatment Effects

The effects documented above could vary across physicians. If the chatbot operates by changing what patients know and ask for, its influence on clinical decisions could depend on how receptive each physician is to patient input and on whether their baseline practice involves overprescribing that treated patients can help constrain. We examine heterogeneity along both dimensions, using measures from the physician survey and from pre-experiment prescribing records.

We first split physicians by their stance toward patient input, using two measures from the physician survey. We asked physicians whether they view medical knowledge acquired by patients from external sources as helpful and whether patients should strictly follow physicians' recommendations. These measures capture variation in physician openness and authority style. Panel A of Table 7 shows that the reduction in medication prescribing and increase in diagnostic testing are substantially larger among physicians who report being highly open to patient-acquired health knowledge. In contrast, Panel B shows that the treatment effects are significantly smaller among physicians who place strong emphasis on physician authority. These patterns suggest that the chatbot moves clinical decisions mainly when physicians are receptive to what patients bring to the visit, and far less when

[29] Patients who did not adopt the chatbot are classified as receiving no advice. Patients in the *No Advice* category behave much like the *Control* group, with coefficients far smaller in magnitude than those in the other three groups that receive advice from AI on the topic.

physicians emphasize their own authority.[30]

We next ask whether the effects are concentrated where baseline practice is most intensive.[31] Table 8 examines whether the treatment effects are larger among physicians who express more favorable views toward prescribing specific medications. In the physician survey, we elicit attitudes toward the use of Traditional Chinese Medicine, antibiotics, and opioids, and classify physicians according to whether they report more favorable or more conservative views toward each category. In column (1) of Table 8, the interaction between treatment assignment and a more pro-TCM attitude is negative and statistically significant, indicating that the decline in TCM prescribing is substantially larger among physicians who are more inclined to prescribe TCM. A similar, though imprecisely estimated, pattern appears for antibiotics in column (2). We find no differential effects for opioids, consistent with their low baseline prescribing rates in China.

We then turn to physicians' revealed practice patterns prior to the experiment. Specifically, we classify physicians according to whether their pre-experiment prescribing rates for TCM, antibiotics, and opioids, and their pre-experiment diagnostic testing rates, exceed the median within their department.[32] This measure captures baseline treatment and testing intensity based on actual practice patterns rather than self-reported attitudes. Table 9 reports the results. The reduction in medication use is significantly larger among physicians with higher baseline prescribing intensity. In particular, the decline in TCM prescribing is concentrated among physicians whose pre-experiment TCM prescribing rates were above the departmental median. A similar pattern appears for antibiotics, while we find no differential effects for opioids, again consistent with their low baseline prescribing rates in China. Column (4) turns to diagnostic testing. In contrast to the prescribing results, the interaction between treatment assignment and above-median baseline testing intensity is small and statistically insignificant, while the main effect is positive and significant. The increase in testing is therefore similar across physicians regardless of how intensively they ordered tests before the

[30]We note one alternative channel for the increase in diagnostic testing. Rather than responding to patient requests for testing, physicians facing better-informed and less deferential patients may order additional tests as a defensive response to perceived legal and dispute risk (e.g., Kessler and McClellan, 1996; Studdert et al., 2005; Carroll et al., 2021; Liu et al., 2023; Fang et al., 2025). Our design cannot separate the two channels directly. That the testing increase is concentrated among physicians receptive to patient input and absent among those emphasizing authority is, however, more consistent with physicians' accommodation of patient requests than with a defensive response.

[31]One potential explanation for heterogeneity in prescribing intensity is financially motivated physician-induced demand (e.g., Gruber and Owings, 1996; Liu et al., 2009; Iizuka, 2012; Lu, 2014; Fang et al., 2021). However, such incentives are limited in our institutional setting. Physicians' salaries and bonuses are not tied to drug sales, and hospitals operate under a zero-markup policy for medications (Yi et al., 2015; Yip et al., 2019; Fang et al., 2021).

[32]Rates are measured from January to May 2025, prior to the start of the experiment.

experiment.

The heterogeneity results speak both to the channels through which chatbot access affects clinical practice and to the possible welfare implications. Both the reduction in prescribing and the increase in testing are larger among physicians who report being receptive to patient input. The pattern is consistent with the intervention operating through patients' treatment requests, which could take the form of asking for diagnostic clarification and resisting medication.

The aggregate welfare consequences are difficult to assess, as we do not observe whether individual prescriptions or tests were clinically indicated. The reduction in prescribing, however, is concentrated among physicians whose baseline prescribing was already high relative to their departmental peers, where the scope for reducing clinically unnecessary treatment is largest. Given the documented overprescription in Chinese outpatient care, the reduction in prescribing, particularly for TCM and antibiotics, could be welfare improving. The increase in testing, by contrast, is similar for physicians with high and low baseline testing rates. The additional tests therefore do not appear concentrated among physicians who were testing too little, and the increase may not be welfare improving.

5 Dynamic, Spillover, and Persistence Effects

The effects documented so far arise from patients' direct use of the chatbot during the experiment. We now ask whether they reach further: whether they persist in physician practice after the experiment, spill over to patients without access, or endure for treated patients at later visits.

5.1 Dynamic Effects

To study the temporal dynamics of the intervention, we implement an event-study framework. This approach serves three purposes. First, it provides a diagnostic check on the physician-level randomization by testing for differential pre-trends between *Exposed* and *Unexposed* physicians. Second, it characterizes how outcomes for *Exposed* physicians evolve over the course of the intervention. Third, it allows us to assess whether any effects persist after the formal intervention period ends.

We aggregate the data to the physician-week level and estimate the following specification:

$$Y_{pt} = \alpha + \sum_{k \neq -1} \beta_k \cdot \big(\text{Exposed}_p \times \mathbb{1}\{\text{Week}_t = k\}\big) + \delta_p + \tau_t + \varepsilon_{pt}, \tag{2}$$

where Y_{pt} denotes the outcome for physician p in week t, such as the rate of Traditional Chinese Medicine prescribing or diagnostic testing. The variable Exposed_p equals one for physicians randomized to the *Exposed* group. The indicator $\mathbb{1}\{\text{Week}_t = k\}$ indexes k weeks relative to the intervention start date of June 2, 2025. The coefficient for $k = -1$, corresponding to the week from May 26 to June 1, 2025, is normalized to zero.

The coefficients of interest are $\{\beta_k\}_{k=-12}^{12}$. For $k < -1$, the estimated coefficients capture differences in outcomes between *Exposed* and *Unexposed* physicians prior to the intervention and provide a test for pre-existing differential trends. For $k \geq 0$, the coefficients trace the dynamic response of outcomes for *Exposed* physicians relative to *Unexposed* physicians. The intervention period extends through day 5 of week $k = 4$; coefficients for subsequent weeks therefore capture post-intervention dynamics. Weeks earlier than $k = -12$ and later than $k = +12$ are binned at the respective endpoints. All specifications include physician fixed effects δ_p, which absorb time-invariant differences across physicians, and week fixed effects τ_t, which control for common temporal shocks. Standard errors are clustered at the physician level to account for serial correlation within physicians over time. Physician-week observations are weighted by the number of patient visits.

Figure 3 presents the results for medication prescribing rates (overall, Traditional Chinese Medicine, antibiotics, and opioids), diagnostic testing, and revisit rates. For all outcomes, we observe no evidence of pre-trends prior to the experiment, suggesting balance between *Exposed* and *Unexposed* physicians. Weeks 0 through 3 correspond to the four full weeks of the intervention period, so the coefficients β_0–β_3 capture the treatment effect and its evolution during the experiment. Because the analysis compares outcomes between *Exposed* and *Unexposed* physicians, the estimated effects are mechanically attenuated. Only half of the patients seen by *Exposed* physicians are assigned access to the chatbot, and take-up among eligible patients is incomplete (17.1%). As a result, the magnitudes of the estimates are expected to be small and can be interpreted as roughly one-half of the corresponding intent-to-treat effects and approximately one-tenth of the average treatment-on-the-treated effects.

As shown in Figure 3, we observe clear reductions in overall medication prescribing and Traditional Chinese Medicine prescribing beginning in week 1, while the estimated effects on antibiotic and opioid prescribing are substantially noisier. Starting in week 1, we also observe a clear increase in diagnostic testing rates, along with a noisy and statistically insignificant decrease in revisit rates.

While point estimates vary across weeks, including a null effect in week 0, we find no statistically significant evidence that the treatment effects change between week 1 and week 3. The experiment concludes in the middle of week 4. Beginning in week 4, we find no differences in any of these outcomes between *Exposed* and *Unexposed* physicians. This pattern indicates no persistence in physicians' clinical practices following the intervention. Once the experiment ends, the gap between *Exposed* and *Unexposed* physicians fully closes.

5.2 Within-Physician Spillovers

Having shown that any effect on physician practice disappears once the experiment ends, we next test for within-physician spillovers during the intervention period itself. Specifically, we examine whether interactions with patients who have access to the chatbot affect outcomes for patients without access who see the same physician. To do so, we adopt the event study framework in equation (2) with one modification. We exclude all visits assigned access to the chatbot. This restriction allows us to compare outcomes for patients without chatbot access who see *Exposed* physicians (*Control* group) with outcomes for patients without chatbot access who see *Unexposed* physicians (*Unexposed* group). If within-physician spillovers or rebound effects are present, we would expect to observe differences in outcomes between these two groups of patients, even though neither group has access to the chatbot.

Appendix Figure A3 reports the results for the same set of clinical outcomes. Across all outcomes, we find no statistically significant differences between patients randomly assigned to the *Control* group and those assigned to the *Unexposed* group. This finding is consistent with the patterns observed in Figure 2. Combined with the absence of persistent effects documented in Figure 3, we are able to conclude that interactions with AI-prepared patients do not generate spillovers to other patients. In particular, physicians do not appear to update their clinical practices in ways that affect patients who do not have access to the chatbot, either contemporaneously or after the intervention ends. This pattern suggests that the effects of the chatbot operate primarily through direct patient use rather than through lasting changes in physician behavior.

One potential concern is that exposure to AI-prepared patients may generate spillovers across physicians, for example through informal communication or peer learning. Such spillovers would bias our estimates of persistence and within-physician spillovers toward zero by contaminating

the comparison group if *Unexposed* physicians indirectly update their practice over time. Several features of our setting mitigate this concern. First, physicians conduct consultations independently in separate rooms, limiting opportunities to observe or discuss specific patient interactions. Second, prescribing patterns among *Unexposed* physicians who never interact with treated patients remain stable over time. Third, for across-physician spillovers to drive the within-physician estimates to zero, the magnitude of across-physician learning would need to be comparable to the magnitude of within-physician learning generated by direct exposure to treated patients. We consider across-physician learning of that scale unlikely in our setting. Taken together, these patterns indicate that across-physician spillovers are unlikely to change the qualitative interpretation of our results.

From a policy perspective, these results raise concerns about unequal access and uptake. Because chatbot adopters are disproportionately younger, male, and employed, the benefits of generative AI tools may flow primarily to already advantaged groups, potentially widening existing health disparities. Policies aimed at expanding access, improving usability, or targeting underserved populations may therefore be necessary to ensure that generative AI tools benefit a broader patient population.

5.3 Patient-Level Persistence

The intervention produces no lasting change in physician practice and does not spill over to patients without access. We now ask whether it persists for treated patients at their later visits. We track the same patients over October to December 2025, three to six months after the intervention ended.[33] We first ask whether access to the chatbot affects patients' subsequent outpatient utilization, measured by whether they have any visit and by their number of visits. We then ask whether, for patients with at least one visit, their prescriptions and diagnostic tests differ.[34]

We begin with subsequent use of outpatient care. Panels (a) and (b) of Figure 4 show that access to the chatbot does not change patients' later use of care. Between October and December 2025, treated patients are slightly less likely to have any visit and make slightly fewer visits than control patients, but neither difference is statistically significant, and Appendix Table A11 confirms these

[33]The chatbot link expires immediately after the patient's visit, so treated patients have no further access during the later period.

[34]A patient may visit more than once during the experiment, with different visits falling into different treatment groups, so we classify each patient by the assignment of their first experimental-period visit.

null results. Although chatbot access marginally reduces two-week revisits, it has no detectable effect on utilization months later, when visits are more likely to be for unrelated conditions. The absence of a difference in utilization also matters for the clinical-practice results that follow. Because treated and control patients have later visits at similar rates, any difference in their prescriptions and tests is less likely to be driven by selection.

We then examine clinical practice at the later visits, reported in Panels (c) and (d) of Figure 4. The reduction in prescribing has faded, and treated and control patients are now equally likely to receive a medication. The increase in diagnostic testing, however, remains. Appendix Table A12 shows that visits by treated patients are still 1.6 percentage points more likely to include a diagnostic test. Because the effect is marginally significant and persists for testing but not for prescribing, we read the evidence as suggestive rather than conclusive. Even so, that the testing effect endures suggests that the chatbot's influence may be long-lasting, whether because patients continue to consult other AI tools on their own or because a single exposure durably shifts the preferences and requests they bring to later consultations.

6 Patient Behavior and Physician–Patient Relationship

Having documented changes in clinical decisions, we now examine whether chatbot access also affects patient behavior and perceptions of the consultation. Figure 5 summarizes these outcomes by treatment status, drawing on both administrative medical records and the patient survey. Panels (a) through (c) use administrative medical records. Panel (a) reports patient attrition, measured as whether patients who scheduled an appointment failed to attend their consultation. We find no evidence that access to the chatbot discourages patients from seeing physicians or substitutes for in-person care, as attendance rates are similar across the *Treatment*, *Control*, and *Unexposed* groups. Panels (b) and (c) capture realized post-consultation behavior, measuring whether patients purchased prescribed medications and whether they completed prescribed diagnostic tests following the visit. For both outcomes, we find no statistically meaningful differences across the *Treatment*, *Control*, and *Unexposed* groups. This suggests that access to the chatbot does not materially affect patients' decisions to follow through on physicians' recommendations along these observable dimensions.

Panels (d) through (f) are based on patient survey responses and therefore reflect a more selected

sample. While these survey-based measures should be interpreted with caution, they provide insight into how access to the chatbot may have affected the consultation experience. Treated patients report lower overall satisfaction, with 83.3% of respondents reporting being very satisfied, compared with approximately 86% in both the *Control* and *Unexposed* groups. Treated patients also perceive communication with physicians as less smooth, as reflected in a 2 percentage point reduction in the likelihood of rating the interaction as very smooth. Finally, intended compliance with physicians' recommendations declines substantially among treated patients, with the share reporting an intention to fully comply falling from 90% in the *Control* and *Unexposed* groups to 84.5% in the *Treatment* group. Appendix Table A13 confirms these patterns in regression estimates. Patients with chatbot access report lower satisfaction, less smooth communication with physicians, and lower intended compliance, with the communication and compliance effects statistically significant despite the relatively small sample size.

Taken together, these patterns suggest that while the chatbot does not meaningfully alter observable patient behavior, it may negatively affect patients' perceptions of the clinical encounter and their interactions with physicians. One possibility is that when patients act more knowledgeable, they receive less polite responses from physicians, a pattern documented by Currie et al. (2011). Another possibility is that access to additional health information prior to the visit changes patients' expectations about the consultation, including which questions should be addressed or how much explanation they anticipate from the physician. When these expectations are not fully met, patients may report lower satisfaction or perceive communication as less smooth, even if clinical decisions and follow-through remain largely unchanged. This interpretation is consistent with evidence from Ayers et al. (2023), which shows that in response to patient questions posted on social media, AI chatbot replies are rated more favorably than physician responses overall and on measures of empathy and clarity.[35]

While the patient survey provides insight into the consultation experience from the patient's perspective, it reflects a selected sample of respondents. We therefore complement this evidence with a physician survey that captures the consultation from the physicians' perspective. The physician survey achieves a high completion rate and thus does not suffer from the same selection concerns,

[35]The survey additionally asks chatbot users to rate the helpfulness of the AI assistant on a five-point scale, from "not helpful at all" to "very helpful." Among respondents, 84 percent rated the chatbot as very helpful and an additional 11 percent as somewhat helpful, with fewer than 1 percent reporting it was not helpful.

although inference is limited by the smaller sample size of physicians rather than visits. Accordingly, we interpret the physician survey results as exploratory and suggestive.

Figure 6 summarizes physicians' perceptions of patient behavior along four dimensions, comparing responses from physicians in the *Exposed* and *Unexposed* groups. In Panel (a), physicians in the *Exposed* group are more likely to report that patients describe their symptoms clearly, with 24.8% of physicians reporting that patients very clearly describe their symptoms, compared with 19.4% among *Unexposed* physicians. Panel (b) shows a similar pattern for perceived patient understanding of their condition, with 22.9% of *Exposed* physicians reporting great understanding, compared with 19.4% in the *Unexposed* group. Panel (c) indicates smoother perceived physician–patient communication in the *Exposed* group, with 36.7% reporting very smooth communication, relative to 30.6% among *Unexposed* physicians. Taken together, the patient and physician surveys reveal an important divergence in perspectives. From the patient's perspective, access to the chatbot is associated with worse perceived communication and lower satisfaction with the visit, whereas from the physician's perspective, patients appear better prepared and communicate more clearly.

Despite these opposing assessments of communication quality, both patients and physicians report lower adherence or intended compliance with medical advice among treated patients. Panel (d) of Figure 6 shows a 5.7 percentage point reduction in physician-reported patient adherence in the *Exposed* group. This pattern mirrors the patient survey findings and highlights a potential tension between improved patient preparedness and reduced adherence to physicians' recommendations. This is consistent with prior evidence showing that patients with access to expertise exhibit lower adherence (Finkelstein et al., 2022). From a policy perspective, this pattern underscores a potential trade-off: while AI tools may enhance patient understanding and improve aspects of the consultation process, they may also alter authority dynamics in ways that affect adherence, with potential implications for both care delivery and health outcomes.

Appendix Table A14 reports regression estimates. The independent variable, *Share of AI Adopters*, is the fraction of a physician's patients who used the chatbot, which equals zero by construction for *Unexposed* physicians and varies across *Exposed* physicians based on patient take-up. The point estimates are consistent with Figure 6. Physicians with higher patient AI adoption report greater symptom clarity, better patient understanding, smoother communication, and lower patient compliance. However, the estimates are imprecise, with only the clarity and compliance coefficients

reaching marginal statistical significance in Panel B. The small sample size of 171 physicians limits statistical power.

As a final sanity check, we elicit physicians' beliefs about the share of their patients who used generative AI tools prior to consultations during the experimental month. Importantly, physicians were not informed of the experiment or its nature throughout the study period, so these perceptions are formed only through interacting with their patients.[36] Appendix Figure A4 reports the results. *Exposed* physicians report an average perceived patient AI adoption rate of 25.2 percent, compared with 16.2 percent among *Unexposed* physicians, a difference of 9 percentage points consistent with the direction of our experimental design. The perceived rates are considerably higher than the actual take-up recorded in our chatbot usage logs, which is unsurprising given that physician recall is subject to recollection biases and their estimates may also capture patients' use of other AI tools beyond our chatbot. Nonetheless, the gap between *Exposed* and *Unexposed* physicians provides additional confirmation that the treatment assignment meaningfully altered the information environment patients brought into the consultation.

7 Conclusion

This paper provides the first large-scale randomized evidence on how patient access to generative AI affects healthcare delivery in a real-world clinical setting. We show that providing patients with access to a generative AI chatbot prior to outpatient visits meaningfully alters clinical practice. Medication prescribing declines by 4.6 percentage points, with particularly large reductions for Traditional Chinese Medicine and antibiotics, while diagnostic testing increases by 2.7 percentage points, without raising total medical expenditures. We find no evidence of spillovers in physicians' behavior toward patients without chatbot access, indicating that the effects operate primarily through direct patient use rather than persistent changes in physician practice. Because chatbot adopters are disproportionately younger, male, and employed, the absence of spillovers also implies that unequal access to generative AI tools may translate directly into unequal benefits, which could widen existing health disparities.

[36] In our post-visit patient survey, 80.6% of *Treatment*-group respondents who reported using the chatbot also reported telling their physician about it during the consultation. While survey respondents are a selected subset of treated patients, this disclosure rate offers a plausible channel through which physicians inferred patients' chatbot use without being informed of the experiment.

The specific pattern of clinical change we document, reduced prescribing and increased diagnostic testing, mirrors the directional advice we observe in the conversation logs. Consistent with the defensive, liability-driven guardrails encoded in AI training, the chatbot cautions heavily against medications while recommending diagnostic testing freely. Treated patients carry this directionality into their consultations, and part of the effect appears to persist in their later visits months after the experiment ended. More broadly, when AI systems are deployed as consumer-facing tools at scale, the norms encoded in their training propagate into real clinical decisions, with the potential to shift aggregate outcomes such as population-level healthcare utilization.

At the same time, our survey results reveal a tension between the use of generative AI tools and the physician–patient relationship. While physicians report clearer communication and better patient preparation when interacting with AI-assisted patients, patients themselves report lower satisfaction and reduced intended adherence to medical advice. These patterns suggest that generative AI may reshape expectations about clinical encounters and shift the balance of authority within the physician–patient interaction in ways that affect compliance.

Whether these shifts improve welfare is not clear. The reductions in prescribing fall disproportionately on Traditional Chinese Medicine and antibiotics, categories long viewed as overprescribed in China, and are concentrated among physicians whose baseline prescribing was high relative to departmental peers. The added testing may surface conditions that would otherwise be missed. On the other hand, less medication may withhold beneficial treatment, and the added testing imposes costs on patients and the health system. Lower adherence is similarly ambiguous, reflecting either informed skepticism or the disregard of sound advice. Importantly, the direction of these changes is set by the objectives of those who build and govern the AI, which need not coincide with the objectives that optimize healthcare delivery.

Taken together, our findings speak to a broader transformation in expert–client markets. Healthcare is a canonical credence good, but it is far from the only one. Legal services, financial advice, and skilled trades share the same underlying information asymmetry that generative AI is now positioned to reduce. Our results suggest that generative AI does more than narrow that asymmetry. The priorities and cautions embedded in AI training travel with the information into real decisions, introducing new frictions in expert–client relationships along the way. While we document this in healthcare, the same mechanism may operate wherever clients consult generative AI before turning

to an expert. As people increasingly rely on these tools, from health questions to financial decisions, the design choices of a small number of developers can propagate into decision-making across many domains of life.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *The Quarterly Journal of Economics*, 2023, *138* (1), 1–35.

Abaluck, Jason, Robert Pless, Nirmal Ravi, Anja Sautmann, and Aaron Schwartz, "Does LLM Assistance Improve Healthcare Delivery? An Evaluation Using On-site Physicians and Laboratory Tests," *National Bureau of Economic Research Working Paper*, 2026.

Angerer, Silvia, Daniela Glätzle-Rützler, Wanda Mimra, Thomas Rittmannsberger, and Christian Waibel, "The Value of Rating Systems in Credence Goods Markets," *The Economic Journal*, 2026, p. ueag011.

Arrow, Kenneth J, "Uncertainty and the Welfare Economics of Medical Care," *The American Economic Review*, 1963, *53* (5), 941–973.

Artmann, Elisabeth, Hessel Oosterbeek, and Bas Van Der Klaauw, "Do Doctors Improve the Health Care of Their Parents? Evidence from Admission Lotteries," *American Economic Journal: Applied Economics*, 2022, *14* (3), 164–184.

Ash, Elliott, Sergio Galletta, and Giacomo Opocher, "BallotBot: Can AI Strengthen Democracy?," *Available at SSRN*, 2025.

Ayers, John W, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth et al., "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," *JAMA internal medicine*, 2023, *183* (6), 589–596.

Badinski, Ivan, Amy Finkelstein, Matthew Gentzkow, and Peter Hull, "Geographic Variation in Healthcare Utilization: The Role of Physicians," *National Bureau of Economic Research Working Paper*, 2023.

Balafoutas, Loukas, Adrian Beck, Rudolf Kerschbamer, and Matthias Sutter, "What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods," *Review of Economic Studies*, 2013, *80* (3), 876–891.

_ , Rudolf Kerschbamer, and Matthias Sutter, "Second-Degree Moral Hazard in a Real-World Credence Goods Market," *The Economic Journal*, 2017, *127* (599), 1–18.

Bertomeu, Jeremy, Yupeng Lin, Yibin Liu, and Zhenghui Ni, "The Impact of Generative AI on Information Processing: Evidence from the Ban of ChatGPT in Italy," *Journal of Accounting and Economics*, 2025, p. 101782.

Bini, Pietro, Lin William Cong, Xing Huang, and Lawrence J Jin, "Behavioral Economics of AI: LLM Biases and Corrections," *Available at SSRN 5213130*, 2025.

Bowen, Donald E, Luke CD Stein, S McKay Price, and Ke Yang, "Measuring and Mitigating Racial Disparities in LLMs: Evidence from a Mortgage Underwriting Experiment," *Working Paper*, 2025.

Bronnenberg, Bart J, Jean-Pierre Dubé, Matthew Gentzkow, and Jesse M Shapiro, "Do Pharmacists Buy Bayer? Informed Shoppers and the Brand Premium," *The Quarterly Journal of Economics*, 2015, *130* (4), 1669–1726.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond, "Generative AI at Work," *The Quarterly Journal of Economics*, 2025, *140* (2), 889–942.

Busse, Meghan R, Ayelet Israeli, and Florian Zettelmeyer, "Repairing the Damage: The Effect of Price Knowledge and Gender on Auto Repair Price Quotes," *Journal of Marketing Research*, 2017, *54* (1), 75–95.

Carroll, Caitlin, David M Cutler, and Anupam Jena, "How Do Physicians Respond to Malpractice Allegations? Evidence from Florida Emergency Departments," *National Bureau of Economic Research Working Paper*, 2021.

Chan, Alex, "Optimal Medical Liability for AI," *National Bureau of Economic Research Working Paper*, 2026.

Chen, Stacey H, Jennjou Chen, Hongwei Chuang, and Tzu-Hsin Lin, "Physicians Treating Physicians: Relational and Informational Advantages in Treatment and Survival," *Journal of Labor Economics*, 2025, *43* (1), 15–46.

Chen, Yiqun, Petra Persson, and Maria Polyakova, "The Roots of Health Inequality and the Value of Intrafamily Expertise," *American Economic Journal: Applied Economics*, 2022, *14* (3), 185–223.

Cheung, Vanessa, Maximilian Maier, and Falk Lieder, "Large Language Models Show Amplified Cognitive Biases in Moral Decision-Making," *Proceedings of the National Academy of Sciences*, 2025, *122* (25), e2412015122.

Choi, Hongseok, Jeongbin Kim, Matthew Kovach, Kyu-Min Lee, Euncheol Shin, and Hector Tzavellas, "Do People Follow AI Advice? Evidence from a Pension Portfolio Choice Experiment," *arXiv preprint arXiv:2608.11371*, 2026.

Choi, Jonathan H, Amy B Monahan, and Daniel Schwarcz, "Lawyering in the Age of Artificial Intelligence," *Minn. L. Rev.*, 2024, *109*, 147.

Choukhmane, Taha, Tim de Silva, Weidong Lin, and Matthew Akuzawa, "AI Financial Advice: Supply, Demand, and Life Cycle Implications," *National Bureau of Economic Research Working Paper*, 2026.

Conlon, John J and Peter Schwardmann, "AI Sycophancy and Decisions," 2026.

Cui, Zheyuan Kevin, 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," *Available at SSRN 4945566*, 2024.

Currie, Janet, W Bentley MacLeod, and Kate Musen, "Doctor Decision-Making and Patient Outcomes," *Journal of Economic Literature*, 2026, *64* (1), 141–194.

_ , Wanchuan Lin, and Juanjuan Meng, "Social Networks and Externalities from Gift Exchange: Evidence from a Field Experiment," *Journal of Public Economics*, 2013, *107*, 19–30.

_ , _ , and _ , "Addressing Antibiotic Abuse in China: An Experimental Audit Study," *Journal of Development Economics*, 2014, *110*, 39–51.

_ , _ , and Wei Zhang, "Patient Knowledge and Antibiotic Abuse: Evidence from an Audit Study in China," *Journal of Health Economics*, 2011, *30* (5), 933–949.

Cutler, David, Jonathan S Skinner, Ariel Dora Stern, and David Wennberg, "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending," *American Economic Journal: Economic Policy*, 2019, *11* (1), 192–221.

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

Das, Jishnu, Alaka Holla, Aakash Mohpal, and Karthik Muralidharan, "Quality and Accountability in Health Care Delivery: Audit-Study Evidence from Primary Care in India," *American Economic Review*, 2016, *106* (12), 3765–3799.

Dell'Acqua, Fabrizio, Edward McFowland III, 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 Technology & Operations Mgt. Unit Working Paper*, 2023, (24-013).

Dillon, Eleanor W, Sonia Jaffe, Nicole Immorlica, and Christopher T Stanton, "Shifting Work Patterns with Generative AI," *National Bureau of Economic Research Working Paper*, 2025.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite, "Is More Information Better? The Effects of "Report Cards" on Health Care Providers," *Journal of Political Economy*, 2003, *111* (3), 555–588.

Dulleck, Uwe and Rudolf Kerschbamer, "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods," *Journal of Economic literature*, 2006, *44* (1), 5–42.

Dupas, Pascaline, "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya," *American Economic Journal: Applied Economics*, 2011, *3* (1), 1–34.

_ , "Health Behavior in Developing Countries," *Annual Review of Economics*, 2011, *3* (1), 425–449.

Fang, Hanming, Ming Li, and Jia Xiang, "Fear and Risk Perception: Understanding Physicians' Dynamic Responses to Malpractice Lawsuits," *National Bureau of Economic Research Working Paper*, 2025.

_ **, Xiaoyan Lei, Julie Shi, and Xuejie Yi**, "Physician-Induced Demand: Evidence from China's Drug Price Zero-markup Policy," *National Bureau of Economic Research Working Paper*, 2021.

Filippas, Apostolos, John J Horton, and Benjamin S Manning, "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?," in "Proceedings of the 25th ACM Conference on Economics and Computation" 2024, pp. 614–615.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams, "Sources of Geographic Variation in Health Care: Evidence from Patient Migration," *The Quarterly Journal of Economics*, 2016, *131* (4), 1681–1726.

_ **, Petra Persson, Maria Polyakova, and Jesse M Shapiro**, "A Taste of Their Own Medicine: Guideline Adherence and Access to Expertise," *American Economic Review: Insights*, 2022, *4* (4), 507–526.

Fleming-Dutra, Katherine E, Adam L Hersh, Daniel J Shapiro, Monina Bartoces, Eva A Enns, Thomas M File Jr, Jonathan A Finkelstein, Jeffrey S Gerber, David Y Hyun, Jeffrey A Linder et al., "Prevalence of Inappropriate Antibiotic Prescriptions among US Ambulatory Care Visits, 2010-2011," *JAMA*, 2016, *315* (17), 1864–1873.

Fong, Yuk-fai, Ting Liu, and Xiaoxuan Meng, "Trust Building in Credence Goods Markets," *American Economic Journal: Microeconomics*, 2022, *14* (1), 490–528.

Frakes, Michael, Jonathan Gruber, and Anupam Jena, "Is Great Information Good Enough? Evidence from Physicians as Patients," *Journal of Health Economics*, 2021, *75*, 102406.

Garg, Prashant and Thiemo Fetzer, "Artificial Intelligence Health Advice Accuracy Varies across Languages and Contexts," *arXiv preprint arXiv:2504.18310*, 2025.

Gottschalk, Felix, Wanda Mimra, and Christian Waibel, "Health Services as Credence Goods: A Field Experiment," *The Economic Journal*, 2020, *130* (629), 1346–1383.

Gruber, Jonathan and Maria Owings, "Physician Financial Incentives and Cesarean Section Delivery," *RAND Journal of Economics*, 1996, *27* (1).

Grytten, Jostein, Irene Skau, and Rune Sørensen, "Do Expert Patients Get Better Treatment than Others? Agency Discrimination and Statistical Discrimination in Obstetrics," *Journal of Health Economics*, 2011, *30* (1), 163–180.

He, Haoran and Jingwen Xia, "Helping Doctors' AI-Assisted Health Care Decision-Making: The Role of Social Norms," *Journal of Economic Behavior & Organization*, 2026, *246*, 107597.

Holz, Justin E., Ricardo Perez-Truglia, Andrew Simon, and Alejandro Zentner, "Taxpayer Behavior in the Age of AI: A Field Experiment on Property Tax Appeals," *National Bureau of Economic Research*, 2026.

Iizuka, Toshiaki, "Physician Agency and Adoption of Generic Pharmaceuticals," *American Economic Review*, 2012, *102* (6), 2826–2858.

Johnson, Erin M and M Marit Rehavi, "Physicians Treating Physicians: Information and Incentives in Childbirth," *American Economic Journal: Economic Policy*, 2016, *8* (1), 115–141.

Kaiser Family Foundation, "Poll: Most Who Use Artificial Intelligence Doubt AI Chatbots Provide Accurate Health Information," Technical Report, Kaiser Family Foundation 2024.

Kerschbamer, Rudolf, Daniel Neururer, and Matthias Sutter, "Credence Goods Markets, Online Information and Repair Prices: A Natural Field Experiment," *Journal of Public Economics*, 2023, *222*, 104891.

Kessler, Daniel and Mark McClellan, "Do Doctors Practice Defensive Medicine?," *The Quarterly Journal of Economics*, 1996, *111* (2), 353–390.

Kolstad, Jonathan T, "Information and Quality when Motivation Is Intrinsic: Evidence from Surgeon Report Cards," *American Economic Review*, 2013, *103* (7), 2875–2910.

Kotwani, Anita, Chand Wattal, Shashi Katewa, PC Joshi, and Kathleen Holloway, "Factors Influencing Primary Care Physicians to Prescribe Antibiotics in Delhi India," *Family Practice*, 2010, *27* (6), 684–690.

Kravitz, Richard L, Ronald M Epstein, Mitchell D Feldman, Carol E Franz, Rahman Azari, Michael S Wilkes, Ladson Hinton, and Peter Franks, "Influence of Patients' Requests for Direct-to-Consumer Advertised Antidepressants: A Randomized Controlled Trial," *JAMA*, 2005, *293* (16), 1995–2002.

Levitt, Steven D and Chad Syverson, "Market Distortions when Agents are Better Informed: The Value of Information in Real Estate Transactions," *The Review of Economics and Statistics*, 2008, *90* (4), 599–611.

Li, Hanzhe, Jin Li, Ye Luo, and Xiaowei Zhang, "AI Persuasion, Bayesian Attribution, and Career Concerns of Doctors," *arXiv preprint arXiv:2410.01114*, 2024.

Li, Yongbin, Jing Xu, Fang Wang, Bin Wang, Liqun Liu, Wanli Hou, Hong Fan, Yeqing Tong, Juan Zhang, and Zuxun Lu, "Overprescribing in China, Driven by Financial Incentives, Results in Very High Use of Antibiotics, Injections, and Corticosteroids," *Health affairs*, 2012, *31* (5), 1075–1082.

Linder, Jeffrey A, Jason N Doctor, Mark W Friedberg, Harry Reyes Nieva, Caroline Birks, Daniella Meeker, and Craig R Fox, "Time of Day and the Decision to Prescribe Antibiotics," *JAMA Internal Medicine*, 2014, *174* (12), 2029–2031.

Liu, Gordon, Junjian Yi, Ye Yuan, and Shaoyang Zhao, "The Short-and Long-Run Effects of Medical Malpractice Lawsuits on Medical Spending and Hospital Operations in China," *Journal of Comparative Economics*, 2023, *51* (4), 1142–1161.

Liu, Ya-Ming, Yea-Huei Kao Yang, and Chee-Ruey Hsieh, "Financial Incentives and Physicians' Prescription Decisions on the Choice Between Brand-Name and Generic Drugs: Evidence from Taiwan," *Journal of Health Economics*, 2009, *28* (2), 341–349.

Lopez, Carolina, Anja Sautmann, and Simone G Schaner, "Do Patients Value High-Quality Medical Care? Experimental Evidence from Malaria Diagnosis and Treatment," *National Bureau of Economic Research Working Paper*, 2024.

_ , _ , and Simone Schaner, "Does Patient Demand Contribute to the Overuse of Prescription Drugs?," *American Economic Journal: Applied Economics*, 2022, *14* (1), 225–260.

Lu, Fangwen, "Insurance Coverage and Agency Problems in Doctor Prescriptions: Evidence from a Field Experiment in China," *Journal of Development Economics*, 2014, *106*, 156–167.

Luettgau, Lennart, Vanessa Cheung, Magda Dubois, Keno Juechems, Jessica Bergs, Henry Davidson, Bessie O'Dell, Hannah Rose Kirk, Max Rollwage, and Christopher Summerfield, "People Readily Follow Personal Advice from AI but It Does Not Improve Their Well-Being," *arXiv preprint arXiv:2511.15352*, 2025.

Maclean, Johanna Catherine, Justine Mallatt, Christopher J Ruhm, and Kosali Simon, "Economic Studies on the Opioid Crisis: A Review," *National Bureau of Economic Research Working Paper*, 2020.

McGuire, Thomas G, "Physician Agency," *Handbook of Health Economics*, 2000, *1*, 461–536.

Mehmood, Sultan, Christoph Goessmann, and Elliott Ash, "Courts of Tomorrow: Evidence from a Nationwide Rollout of Generative AI," *Available at SSRN 7119447*, 2026.

Mimra, Wanda, Alexander Rasch, and Christian Waibel, "Second Opinions in Markets for Expert Services: Experimental Evidence," *Journal of Economic Behavior & Organization*, 2016, *131*, 106–125.

Mullainathan, Sendhil, Markus Noeth, and Antoinette Schoar, "The Market for Financial Advice: An Audit Study," *National Bureau of Economic Research Working Paper*, 2012.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381* (6654), 187–192.

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.

Rasch, Alexander and Christian Waibel, "What Drives Fraud in a Credence Goods Market?–Evidence from a Field Study," *Oxford Bulletin of Economics and Statistics*, 2018, *80* (3), 605–624.

Reimers, Imke and Joel Waldfogel, "AI and the Quantity and Quality of Creative Products: Have LLMs Boosted Creation of Valuable Books?," *National Bureau of Economic Research Working Paper*, 2026.

Reis, Florian, Louis Agha-Mir-Salim, Richard Hickstein, Moritz Reis, Sophie K Piper, Felix Balzer, and Sebastian Daniel Boie, "Disclaimers and Referral Patterns for Medical Advice across Urgency Levels: Large Language Model Evaluation Study," *Journal of Medical Internet Research*, 2026, *28*, e84668.

Ross, Jillian, Yoon Kim, and Andrew W Lo, "LLM Economicus? Mapping the Behavioral Biases of LLMs via Utility Theory," *arXiv preprint arXiv:2408.02784*, 2024.

Schneider, Henry S, "Agency Problems and Reputation in Expert Services: Evidence from Auto Repair," *The Journal of Industrial Economics*, 2012, *60* (3), 406–433.

Shah, Anand V. and Joshua Y. Levy, "Access to Justice in the Age of AI: Evidence from U.S. Federal Courts," 2026. Working paper.

Sharma, Sonali, Ahmed M Alaa, and Roxana Daneshjou, "A Longitudinal Analysis of Declining Medical Safety Messaging in Generative AI Models," *npj Digital Medicine*, 2025, *8* (1), 592.

Spatharioti, Sofia Eleni, David M Rothschild, Daniel G Goldstein, and Jake M Hofman, "Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment," *arXiv preprint arXiv:2307.03744*, 2023.

Stacey, Dawn, France Légaré, Krystina Lewis, Michael J Barry, Carol L Bennett, Karen B Eden, Margaret Holmes-Rovner, Hilary Llewellyn-Thomas, Anne Lyddiatt, Richard Thomson et al., "Decision Aids for People Facing Health Treatment or Screening Decisions," *Cochrane Database of Systematic Reviews*, 2017, (4).

_ , Krystina B. Lewis, Maureen Smith, Meg Carley, Robert Volk, Elisa E. Douglas, Lissa Pacheco-Brousseau, Jeanette Finderup, Janet Gunderson, Michael J. Barry, Carol L. Bennett, Paulina Bravo, Karina Steffensen, Amédé Gogovor, Ian D. Graham, Shannon E. Kelly, France Légaré, Henning Sondergaard, Richard Thomson, Logan Trenaman, and Lyndal Trevena, "Decision aids for people facing health treatment or screening decisions," *Cochrane Database of Systematic Reviews*, 2024, *2024* (1), CD001431. Art. No.: CD001431. Update of: Cochrane Database Syst Rev. 2017;4:CD001431.

Studdert, David M, Michelle M Mello, William M Sage, Catherine M DesRoches, Jordon Peugh, Kinga Zapert, and Troyen A Brennan, "Defensive Medicine among High-Risk Specialist Physicians in a Volatile Malpractice Environment," *JAMA*, 2005, *293* (21), 2609–2617.

Tao, Xinge, Shuya Zhou, Kai Ding, Sairan Li, Yanzeng Li, Boyou Wu, Qirui Huang, Wangyue Chen, Muzi Shen, En Meng et al., "An LLM Chatbot to Facilitate Primary-to-Specialist Care Transitions: A Randomized Controlled Trial," *Nature Medicine*, 2026, pp. 1–9.

Waight, Hannah, Eddie Yang, Yin Yuan, Solomon Messing, Margaret E Roberts, Brandon M Stewart, and Joshua A Tucker, "State Media Control Influences Large Language Models," *Nature*, 2026, pp. 1–9.

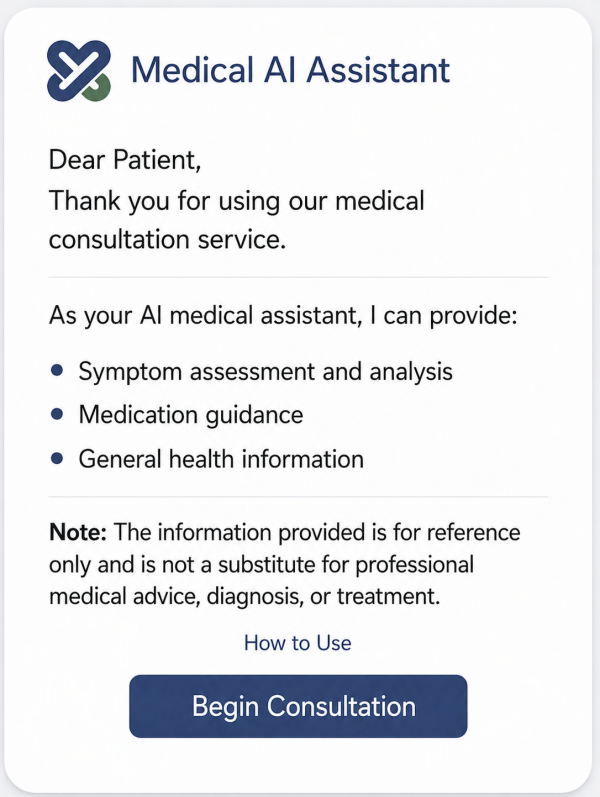
Wan, Peixing, Zigeng Huang, Wenjun Tang, Yulan Nie, Dajun Pei, Shaofen Deng, Jing Chen, Yizhi Zhou, Hongru Duan, Qingyu Chen et al., "Outpatient Reception via Collaboration between Nurses and a Large Language Model: a Randomized Controlled Trial," *Nature Medicine*, 2024, *30* (10), 2878–2885.

Yi, Hongmei, Grant Miller, Linxiu Zhang, Shaoping Li, and Scott Rozelle, "Intended and Unintended Consequences of China's Zero Markup Drug Policy," *Health Affairs*, 2015, *34* (8), 1391–1398.

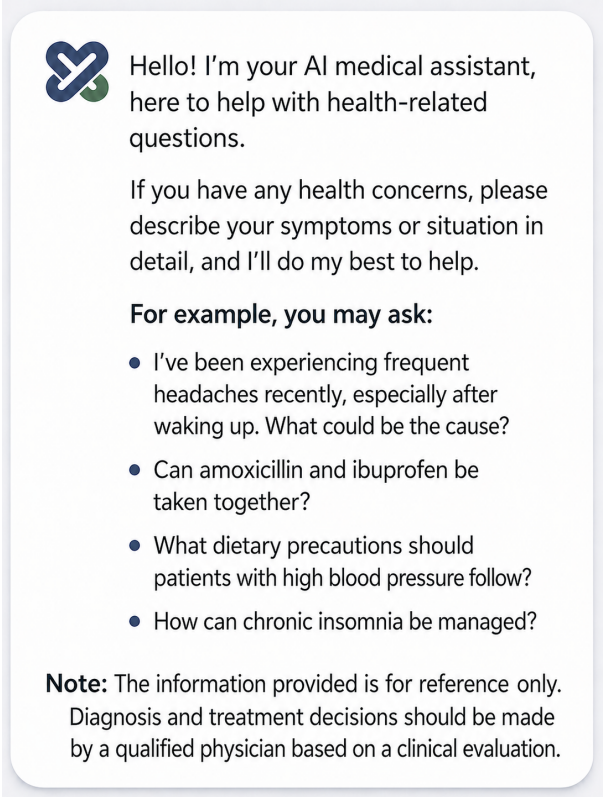
Yip, Winnie, Hongqiao Fu, Angela T Chen, Tiemin Zhai, Weiyan Jian, Roman Xu, Jay Pan, Min Hu, Zhongliang Zhou, Qiulin Chen et al., "10 Years of Health-Care Reform in China: Progress and Gaps in Universal Health Coverage," *The Lancet*, 2019, *394* (10204), 1192–1204.

Figures and Tables

(a) Initial Interface



(b) Consultation Interface



(c) Experimental Design

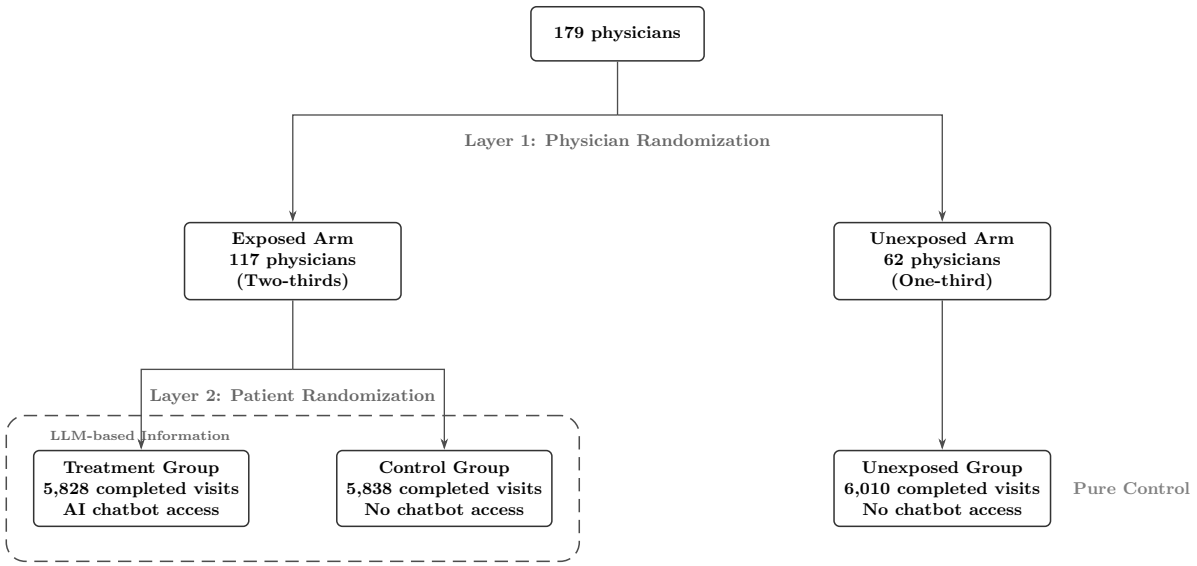


Figure 1. Medical AI Chatbot User Interfaces and Experimental Design Flowchart

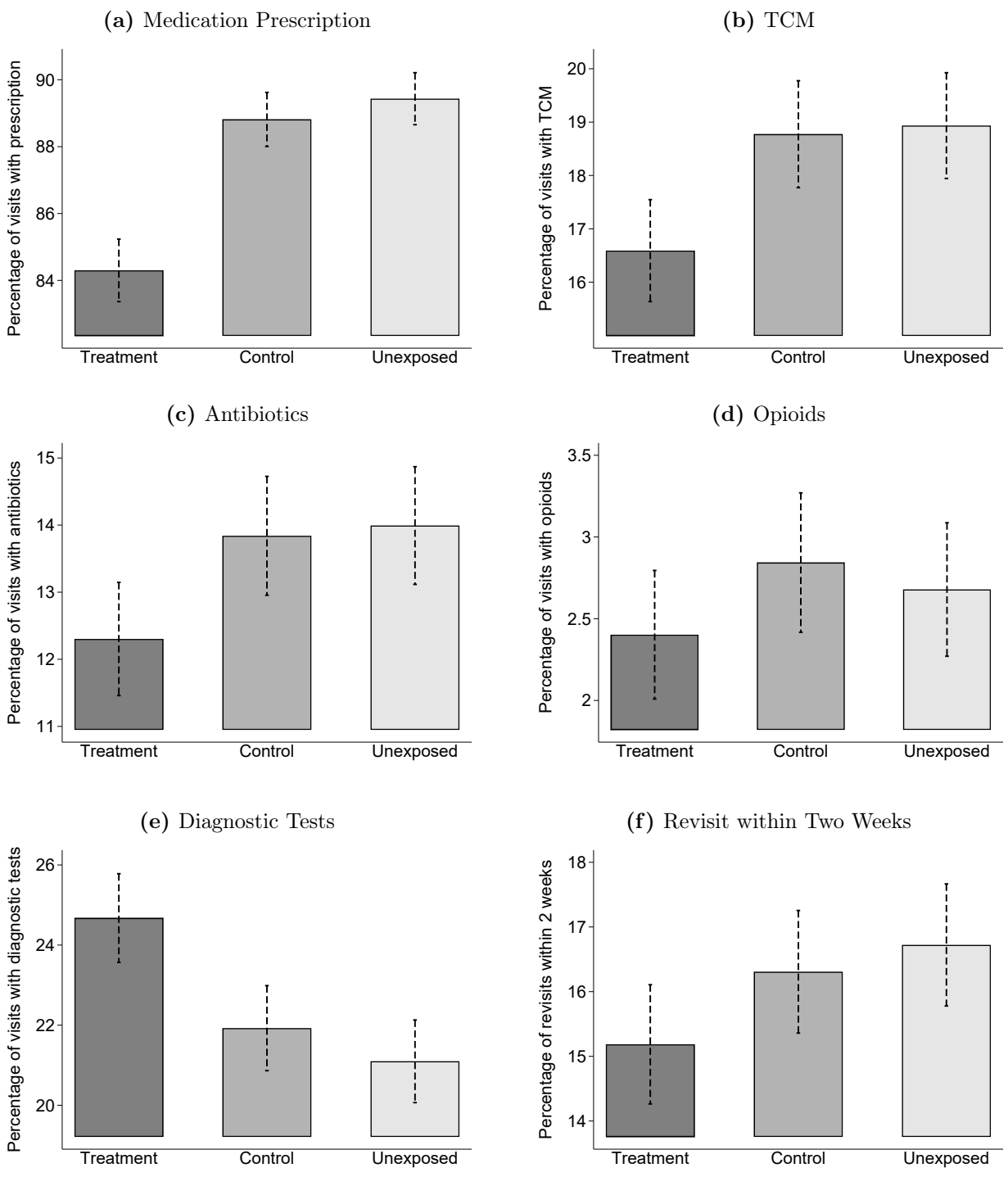
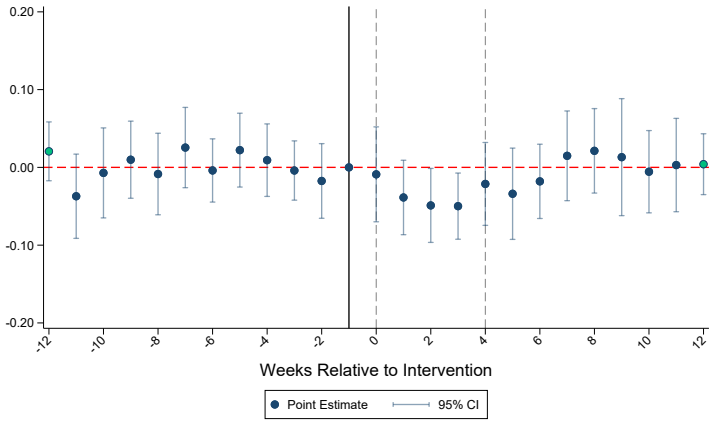


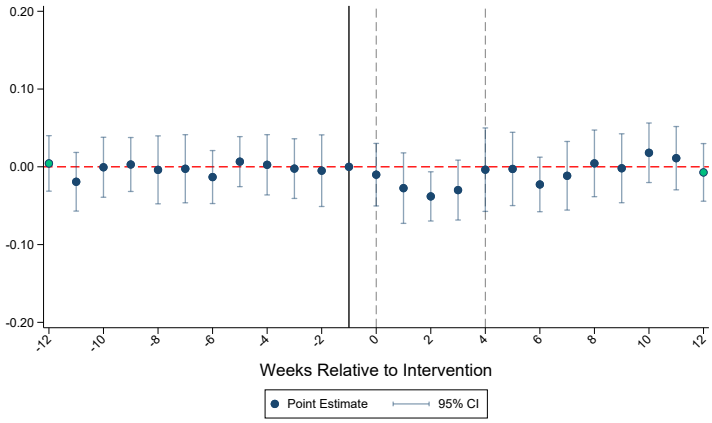
Figure 2. Prescriptions and Clinical Practices by Treatment Status

Notes: This figure reports mean clinical outcomes for patients in the *Treatment*, *Control*, and *Unexposed* groups, with 95% confidence intervals (dashed lines).

(a) Medication Prescription



(b) TCM



(c) Antibiotics

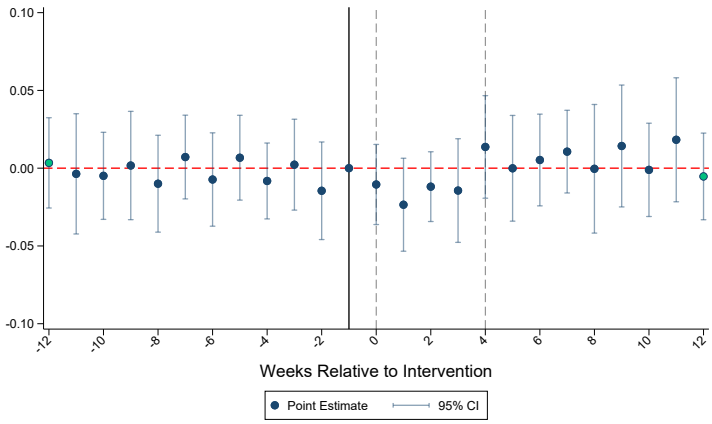
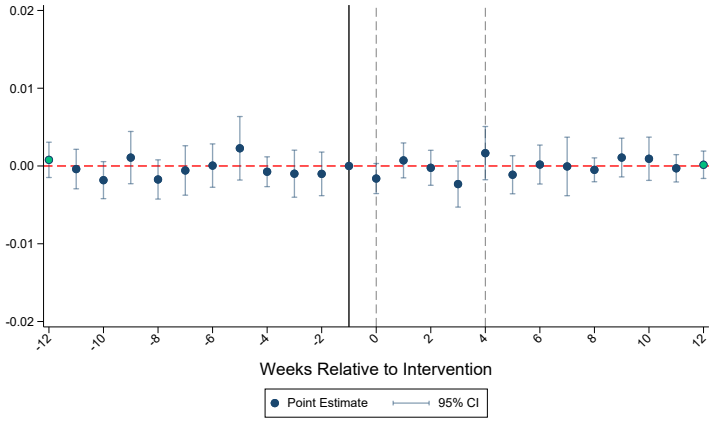
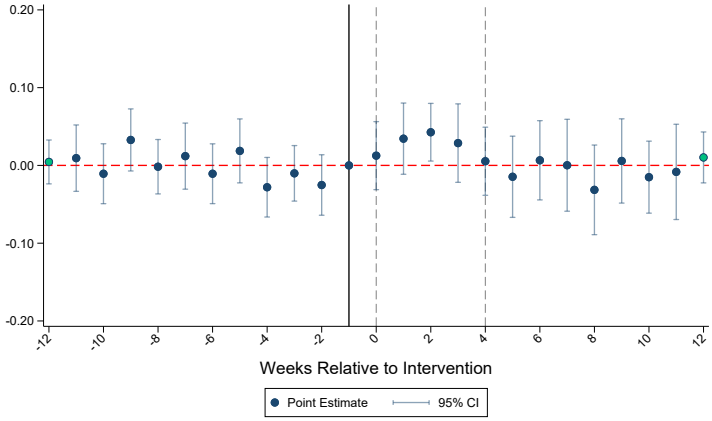


Figure 3. Dynamic Effects of Exposure to Treated Patients on Prescriptions and Clinical Practices

(d) Opioids



(e) Diagnostic Tests



(f) Revisit within Two Weeks

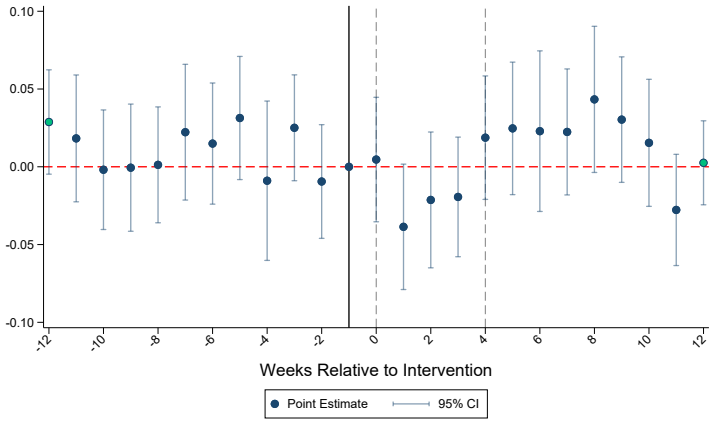


Figure 3. Dynamic Effects of Exposure to Treated Patients on Prescriptions and Clinical Practices (Continued)

Notes: This figure presents the point estimates with 95% confidence intervals from equation (2). Data are aggregated to the physician-week level using all patient visits from 2025, weighted by the number of visits per physician per week. Relative weeks beyond −12 and +12 are grouped at the endpoints. The baseline period is Week −1 (May 26–June 1, 2025). Standard errors are clustered at the physician level.

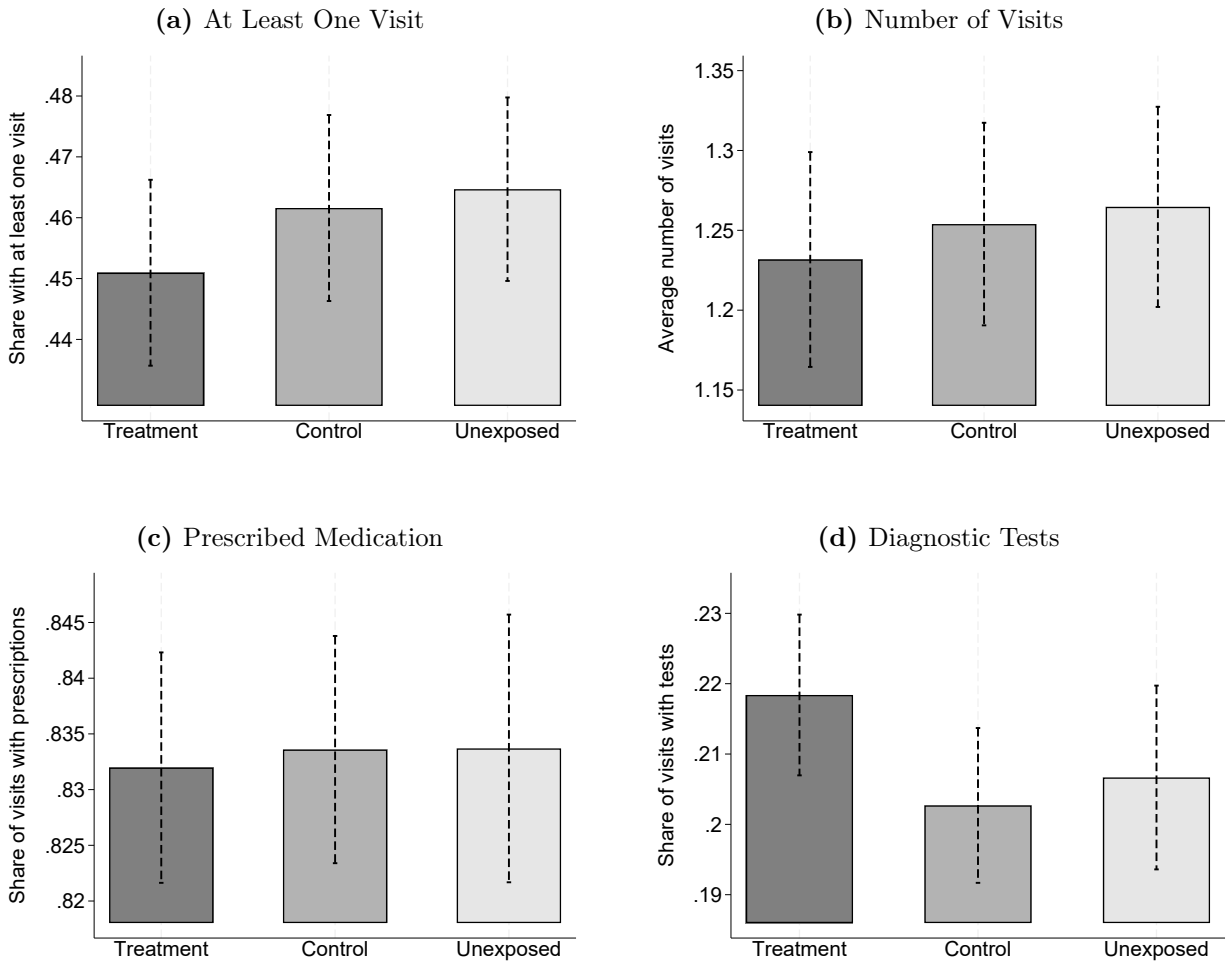


Figure 4. Persistent Effects of Treatment Assignment on Outpatient Utilization and Clinical Practice in the Post-Experiment Period (October–December 2025)

Notes: This figure examines whether assignment to chatbot access during the experiment continues to affect outpatient utilization and clinical practice after the experiment ended. The experiment ran from June 2 to July 3, 2025. We track the same patients from October to December 2025 by matching their subsequent outpatient visits through unique patient identifiers. Because a patient may appear in more than one experimental-period visit with differing assignments, we group each patient by their first such visit. Among patients whose first visit was with an *Exposed* physician, those whose first visit received chatbot access form the *Treatment* group and the remainder form the *Control* group, while patients whose first visit was with an *Unexposed* physician form the *Unexposed* group. Panels (a) and (b) report patient-level outcomes computed for every patient in each group. Panel (a) is the share with at least one outpatient visit between October and December 2025, and Panel (b) is the average number of visits. Panels (c) and (d) report visit-level outcomes among patients with at least one outpatient visit between October and December 2025. Panel (c) is the share of visits with a prescribed medication, and Panel (d) is the share with a diagnostic test. Bars denote group means, and dashed lines denote 95% confidence intervals.

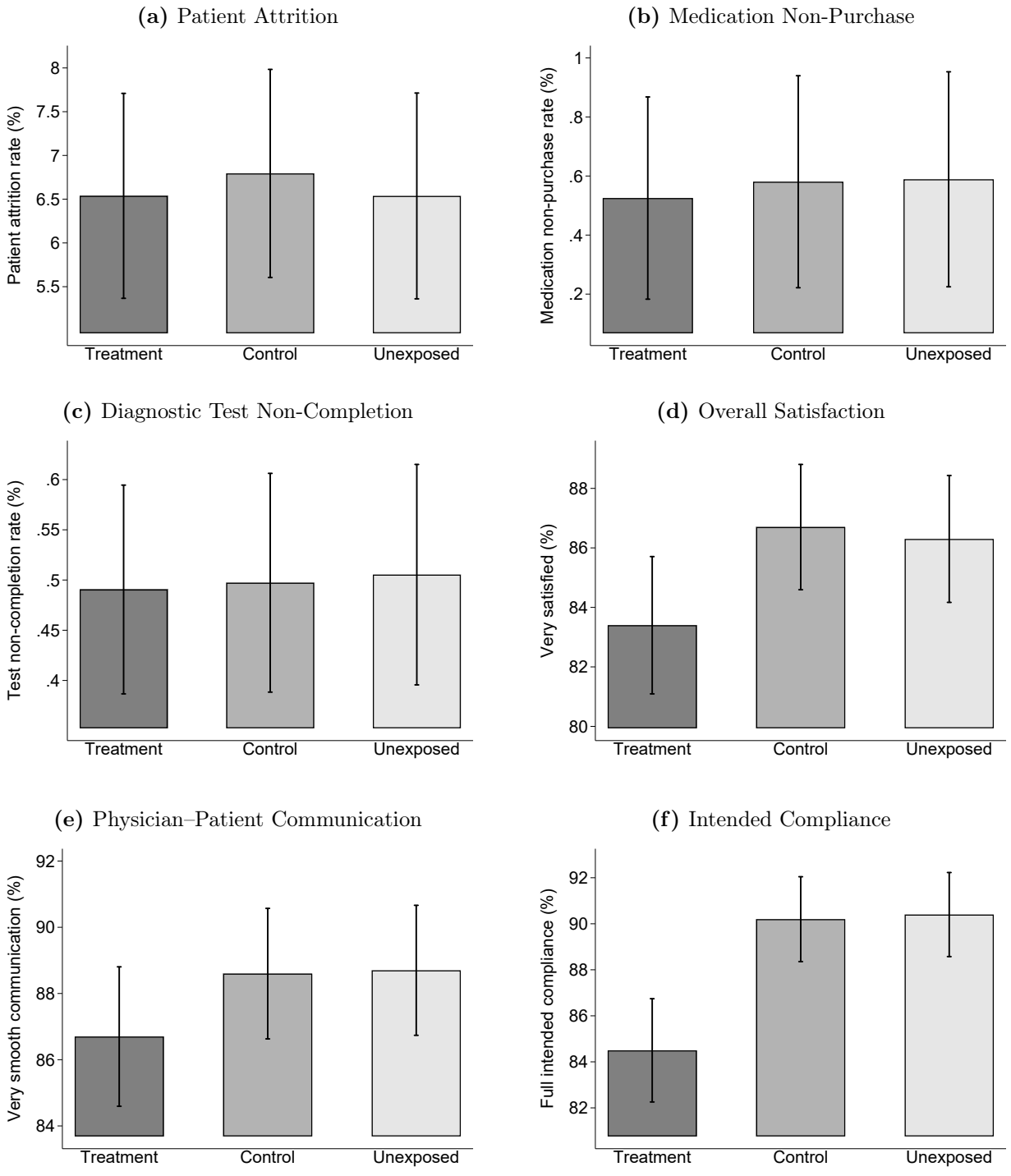


Figure 5. Patient Behavior and Perceptions by Treatment Status

Notes: Panels (a)–(c) report outcomes constructed from administrative medical records and capture patient attrition and realized post-consultation behavior. Panels (d)–(f) report outcomes from a post-consultation patient survey, comparing respondents in the *Treatment*, *Control*, and *Unexposed* groups. Percentages denote the share of patients with the indicated outcome or selecting the indicated response category, with 95% confidence intervals.

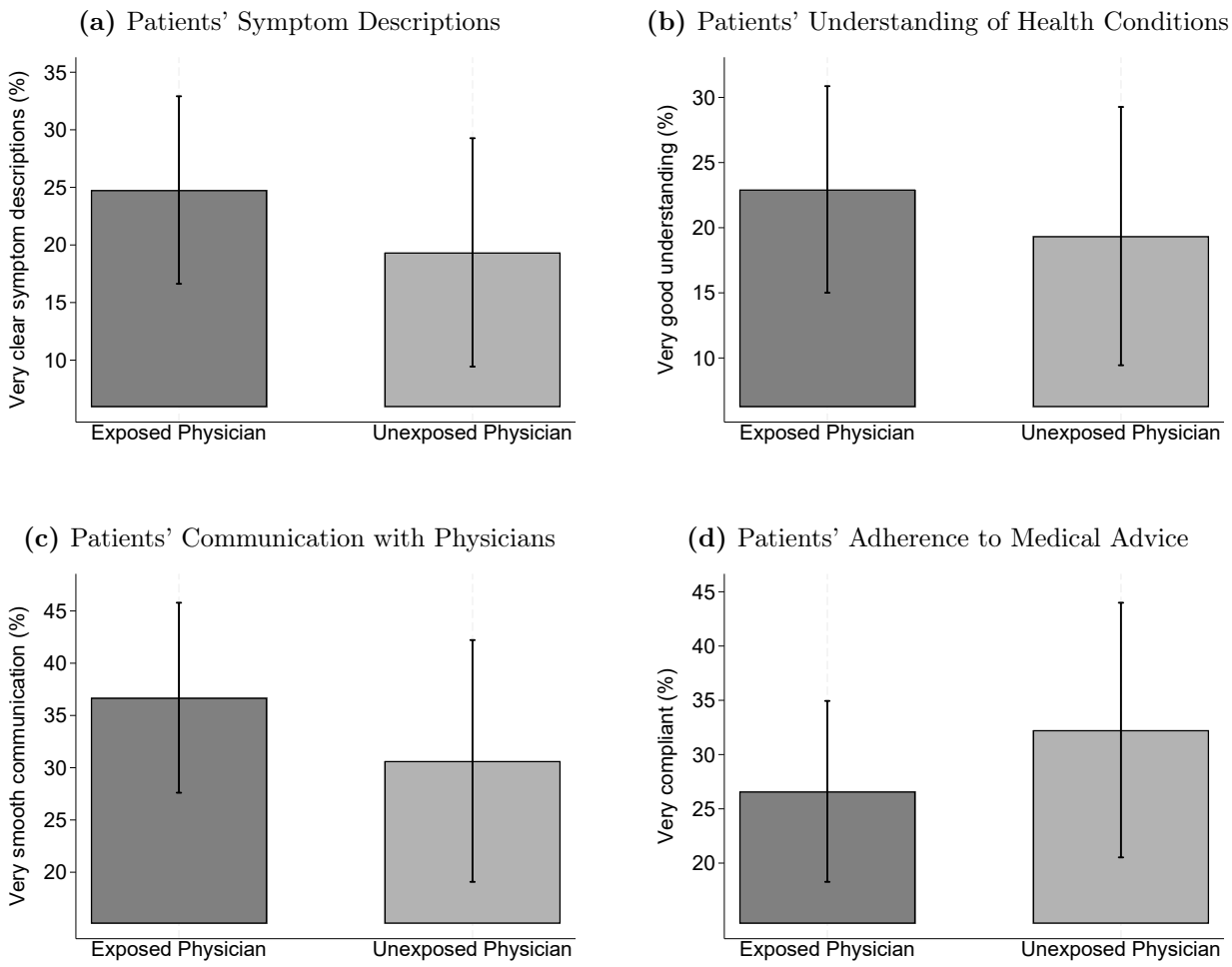


Figure 6. Physicians' Perceptions of Patient Behaviors during Consultations

Notes: This figure reports physicians' survey responses collected after the conclusion of the experiment, comparing physicians randomized to the *Exposed* and *Unexposed* groups. Outcomes reflect physicians' perceptions of patient behavior during consultations. Percentages denote the share of physicians selecting the indicated response category, with 95% confidence intervals.

Table 1. Balance Test: Physician and Patient Characteristics

Panel A: Physician Characteristics						
Variables	Exposed Arm		Unexposed Arm		Diff	p-value
	Mean	S.D.	Mean	S.D.		
Age	37.94	8.52	39.48	9.61	−1.54	0.29
Male	0.61	0.49	0.69	0.46	−0.09	0.25
Senior Physician	0.30	0.46	0.29	0.46	0.01	0.90
College Degree or Above	0.91	0.29	0.92	0.27	−0.01	0.76
Specialty: Internal Medicine	0.34	0.48	0.34	0.48	−0.00	0.97
Specialty: Surgery	0.26	0.44	0.23	0.42	0.04	0.56
Specialty: Psychiatry	0.24	0.43	0.27	0.45	−0.04	0.62
Specialty: Emergency	0.05	0.22	0.08	0.27	−0.03	0.47
Specialty: Other Dept.	0.12	0.33	0.11	0.32	0.01	0.89
Num of Physicians	117		62			

Panel B: Patient Characteristics										
	Treatment		Control		Unexposed		Treat vs Ctrl		Treat vs Unexp	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Diff	p-value	Diff	p-value
Age	45.36	16.92	45.33	16.93	45.49	17.24	0.02	0.95	−0.15	0.63
Male	0.36	0.48	0.36	0.48	0.36	0.48	−0.00	0.96	0.00	0.93
Occupation: Farmer	0.44	0.50	0.44	0.50	0.44	0.50	0.00	0.98	−0.00	0.97
Occupation: Formally Employed	0.24	0.43	0.23	0.42	0.24	0.43	0.01	0.20	−0.01	0.18
Occupation: Student	0.09	0.29	0.09	0.29	0.09	0.29	−0.00	0.92	0.00	0.89
Occupation: Other	0.07	0.26	0.07	0.26	0.07	0.26	−0.00	0.72	0.00	0.69
Unemployed or Retired	0.16	0.36	0.17	0.37	0.16	0.37	−0.01	0.23	−0.00	0.82
First-time Visit	0.70	0.46	0.70	0.46	0.70	0.46	−0.00	0.95	0.00	0.95
Advance Booking Time (hours)	24.63	27.08	24.90	28.02	24.96	27.59	-0.28	0.55	-0.32	0.48
Num of Visits	5,828		5,838		6,010					

Table 2. Patient–AI Conversation Patterns by Topic

	Mention Rate		AI Stance (% of AI mentions)			
Topic	Patient	AI	Rec Only	Rec+Caution	Caution Only	Neither
Diagnostic testing	8.5%	58.2%	94.5%	3.0%	0.3%	2.2%
General medication	11.4%	58.9%	19.1%	44.2%	25.5%	11.3%
Traditional Chinese Medicine	0.8%	11.0%	3.8%	3.8%	87.0%	5.3%
Antibiotics	0.1%	18.2%	7.8%	13.4%	74.2%	4.6%

Notes: The sample consists of 1,192 conversation turns from 956 visits with chatbot access. Mention rates are shares of all turns in which the topic appears, identified through keyword matching at the sentence level. Patient mentions are based on the patient message; AI mentions are based on the chatbot's reply. The four AI attitude categories are constructed from two independent binary flags assigned by GPT-4o for each turn in which the AI mentions the topic: whether the AI recommends the treatment or test, and whether it raises any caution or advises against its use. Attitude shares are computed as percentages of turns in which the AI mentions the topic and therefore sum to 100% within each row. A single turn can mention multiple topics, in which case it is classified independently for each topic.

Table 3. Within-Physician Intent-to-Treat Effects on Clinical Practice Patterns

	Dependent Variables					
	Prescribed medication	TCM	Antibiotics	Opioids	Diagnostic tests	Revisit
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: No Controls						
Treated	-0.045^{***}	-0.022^{***}	-0.015^{**}	-0.004	0.027^{***}	-0.011^{*}
	(0.006)	(0.007)	(0.006)	(0.003)	(0.008)	(0.007)
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
Panel B: Patient Characteristics & Physician FE						
Treated	-0.046^{***}	-0.022^{***}	-0.015^{**}	-0.004	0.027^{***}	-0.012^{*}
	(0.006)	(0.007)	(0.006)	(0.003)	(0.008)	(0.007)
Patient Char.	Y	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666

Notes: This table reports within-physician intent-to-treat estimates of the effect of chatbot access on clinical practice patterns. The sample is restricted to outpatient visits with *Exposed* physicians. Panel A reports estimates without controls; Panel B adds patient characteristics (age, gender, occupation, and an indicator for first visit) and physician fixed effects. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4. Within-Physician Intent-to-Treat Effects on Healthcare Expenditures

	Dependent Variables				
	TCM costs (RMB)	WM costs (RMB)	Diagnostic costs (RMB)	Other costs (RMB)	Total costs (RMB)
	(1)	(2)	(3)	(4)	(5)
Panel A: No Controls					
Treated	−1.686*	−5.276*	6.542	0.174	−0.245
	(0.944)	(2.939)	(5.662)	(3.293)	(7.264)
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666
Panel B: Patient Characteristics & Physician FE					
Treated	−1.762*	−5.021*	6.313	−0.002	−0.472
	(0.945)	(2.952)	(5.622)	(3.388)	(7.279)
Patient Char.	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666

Notes: This table reports within-physician intent-to-treat estimates of the effect of chatbot access on expenditures. The sample is restricted to outpatient visits with *Exposed* physicians. Panel A reports estimates without controls; Panel B adds patient characteristics (age, gender, occupation, and an indicator for first visit) and physician fixed effects. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5. Instrumental Variable Estimates: Treatment Effects on Clinical Practice Patterns

	Dependent Variables					
	Prescribed medication	TCM	Antibiotics	Opioids	Diagnostic tests	Revisit
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: No Controls						
Chatbot Usage	−0.265***	−0.128***	−0.090**	−0.026	0.161***	−0.066*
	(0.037)	(0.041)	(0.037)	(0.017)	(0.046)	(0.039)
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
F-statistic	1198.188					
Panel B: Patient Characteristics & Physician FE						
Chatbot Usage	−0.267***	−0.129***	−0.089**	−0.025	0.161***	−0.070*
	(0.037)	(0.042)	(0.037)	(0.017)	(0.046)	(0.040)
Patient Char.	Y	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
F-statistic	1181.441					

Notes: This table reports within-physician instrumental variable estimates of the effect of chatbot usage on clinical practice patterns, where chatbot usage is instrumented with random assignment to chatbot access. The sample is restricted to outpatient visits with *Exposed* physicians. Panel A reports estimates without controls; Panel B adds patient characteristics (age, gender, occupation, and an indicator for first visit) and physician fixed effects. The reported first-stage F-statistic is the Kleibergen-Paap Wald F-statistic. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6. Instrumental Variable Estimates: Treatment Effects on Healthcare Expenditures

	Dependent Variables				
	TCM costs (RMB) (1)	WM costs (RMB) (2)	Diagnostic costs (RMB) (3)	Other costs (RMB) (4)	Total costs (RMB) (5)
Panel A: No Controls					
Chatbot Usage	−9.884*	−30.933*	38.359	1.023	−1.436
	(5.542)	(17.227)	(33.189)	(19.306)	(42.585)
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666
F-statistic			1198.188		
Panel B: Patient Characteristics & Physician FE					
Chatbot Usage	−10.334*	−29.444*	37.018	−0.009	−2.770
	(5.552)	(17.310)	(32.967)	(19.868)	(42.679)
Patient Char.	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666
F-statistic			1181.441		

Notes: This table reports within-physician instrumental variable estimates of the effect of chatbot usage on healthcare expenditures, where chatbot usage is instrumented with random assignment to chatbot access. The sample is restricted to outpatient visits with *Exposed* physicians. Panel A reports estimates without controls; Panel B adds patient characteristics (age, gender, occupation, and an indicator for first visit) and physician fixed effects. The reported first-stage F-statistic is the Kleibergen-Paap Wald F statistic. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7. Heterogeneous Treatment Effects by Physicians' Attitudes

	Dependent variables					
	Prescribed medication	TCM	Antibiotics	Opioids	Diagnostic tests	Revisit
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Physician Openness						
Treated × Openness	−0.069***	−0.037**	−0.017	−0.000	0.087***	−0.031
	(0.017)	(0.019)	(0.016)	(0.007)	(0.021)	(0.023)
Treated	−0.035***	−0.015	−0.011	0.000	0.007	0.004
	(0.007)	(0.008)	(0.007)	(0.003)	(0.009)	(0.009)
Patient Char.	Y	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
Panel B. Physician Authority						
Treated × Authority	0.036**	0.021*	0.023**	0.004	−0.031	0.007
	(0.016)	(0.012)	(0.011)	(0.007)	(0.020)	(0.017)
Treated	−0.068***	−0.031***	−0.026***	−0.003	0.046***	−0.019
	(0.014)	(0.011)	(0.010)	(0.006)	(0.018)	(0.016)
Patient Char.	Y	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666

Notes: This table reports heterogeneous treatment effects by physicians' attitudes measured in the post-experiment survey. In Panel A, *Openness* equals one if the physician reported that medical knowledge acquired by patients from external sources is "very helpful" to physicians, and zero otherwise. In Panel B, *Authority* equals one if the physician strongly agreed with the statement "Patients should closely follow physicians' recommendations," and zero otherwise. All regressions control for patient characteristics (age, gender, occupation, and first-visit indicator) and include physician fixed effects. Robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 8. Heterogeneous Treatment Effects by Physicians' Attitudes Toward Medications

	Dependent Variables		
	TCM	Antibiotics	Opioids
	(1)	(2)	(3)
Treated × Pro-TCM	−0.029**		
	(0.014)		
Treated × Pro-antibiotic		−0.017	
		(0.012)	
Treated × Pro-opioid			−0.001
			(0.007)
Treated	−0.013	−0.009	0.001
	(0.011)	(0.009)	(0.006)
Patient Char.	Y	Y	Y
Physician FE	Y	Y	Y
Mean	0.18	0.13	0.03
Observations	11,666	11,666	11,666

Notes: This table reports heterogeneous treatment effects by physicians' attitudes toward different medications. Pro-TCM equals one if the physician selected "Equal emphasis on TCM and Western medicine" when asked about the relationship between TCM and Western medicine, and zero if the physician selected "Completely Western medicine" or "Western medicine primary, TCM supplementary." Pro-antibiotic equals one if the physician selected "Use is generally appropriate" when asked about antibiotic use in China, and zero if the physician selected "Obviously overused" or "Slightly overused." Pro-opioid equals one if the physician selected "Use is generally appropriate" when asked about opioid analgesic use in China, and zero otherwise. All regressions control for patient characteristics (age, gender, occupation, and an indicator for first visit) and include physician fixed effects. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9. Heterogeneous Treatment Effects by Physicians' Baseline Practice Intensity

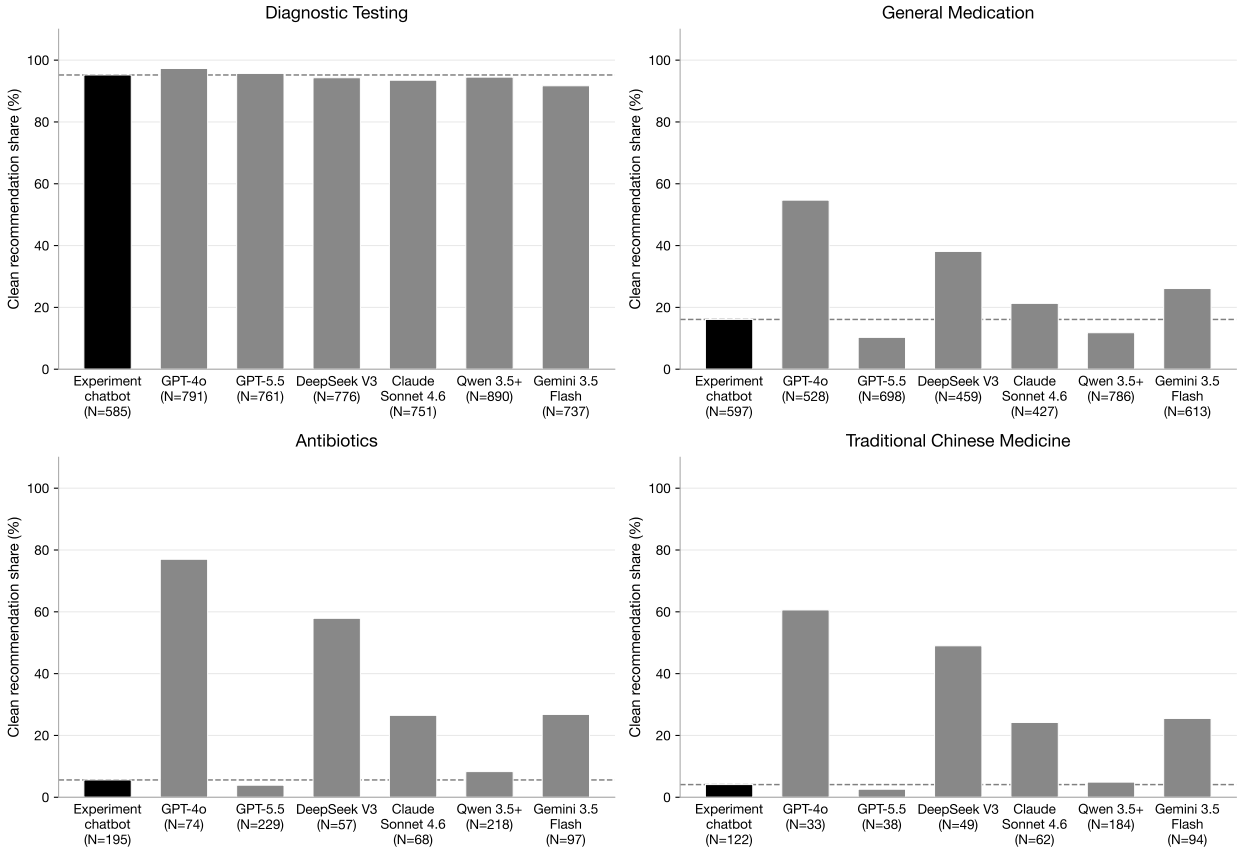
	Dependent Variables			
	TCM	Antibiotics	Opioids	Diagnostic Tests
	(1)	(2)	(3)	(4)
Treated × High-TCM	−0.023**			
	(0.012)			
Treated × High-Antibiotic		−0.019**		
		(0.009)		
Treated × High-Opioid			−0.004	
			(0.004)	
Treated × High-Testing				−0.012
				(0.016)
Treated	−0.005	−0.003	0.004	0.033***
	(0.009)	(0.008)	(0.005)	(0.011)
Patient Char.	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y
Mean	0.18	0.13	0.03	0.23
Observations	11,666	11,666	11,666	11,666

Notes: This table reports heterogeneous treatment effects based on whether a physician has above-median practice intensity within their department prior to the experiment, measured between January and May 2025. High TCM (High Antibiotic, High Opioid) is an indicator equal to one if the physician's pre-experiment mean prescribing rate for TCM (antibiotics, opioids) exceeds the department median, and zero otherwise. High Testing is an indicator equal to one if the physician's pre-experiment mean diagnostic testing rate exceeds the department median, and zero otherwise. All regressions control for patient characteristics (age, gender, occupation, and an indicator for first visit) and include physician fixed effects. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

ONLINE APPENDIX

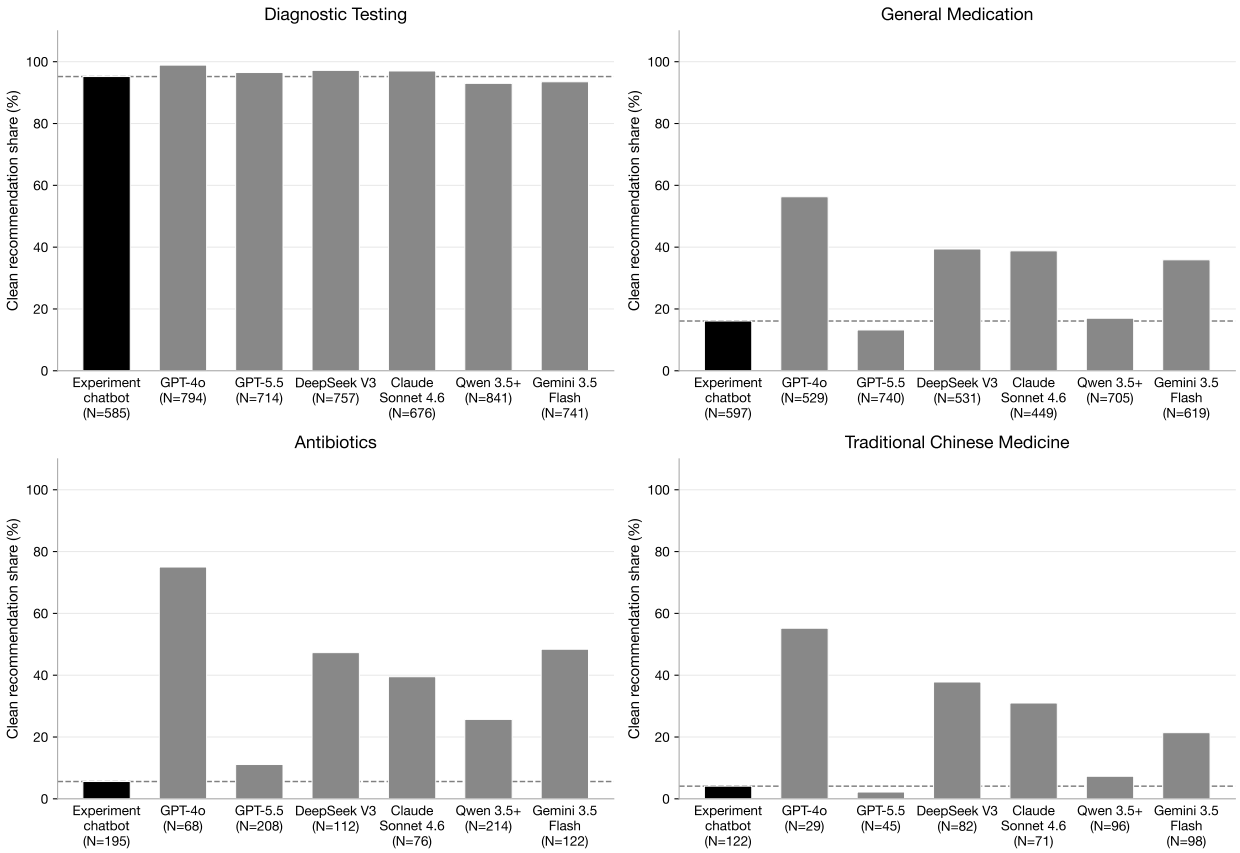
A Additional Figures and Tables

Figure A1. Share of Clean Recommendations by Topic and Model, System Prompt Framing the Model as a Pre-Visit Assistant



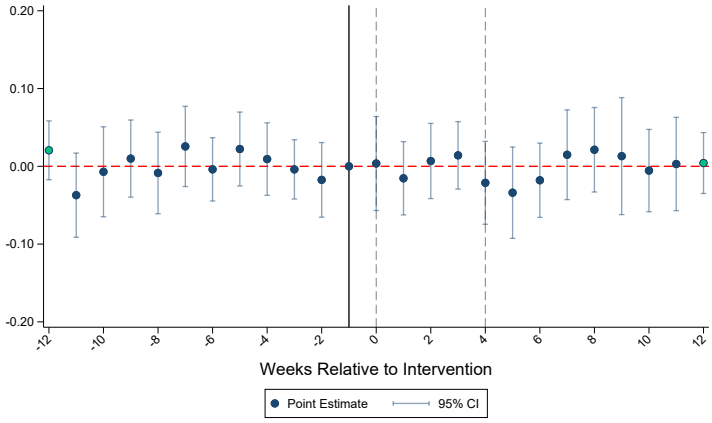
Notes: Each panel corresponds to one of four clinical topics. Bars report the share of topic mentions that the model recommends without any caution (clean recommendations), classified using the stance pipeline described in Appendix B.3. The dark bar is our experiment chatbot; the remaining bars are six comparison models given the same first-turn patient messages under a system prompt framing the model as a pre-visit assistant. The dashed horizontal line marks our experiment chatbot's share in each panel. The number of topic mentions (N) appears below each model. Because the follow-up turns of each conversation cannot be reconstructed for the comparison models, our chatbot is restricted here to its first-turn responses, so its shares differ slightly from Table 2, which uses the full analytic sample. Traditional Chinese Medicine mention counts are small for some models, and those shares should be read with caution.

Figure A2. Share of Clean Recommendations by Topic and Model, System Prompt Framing the Model as a Direct Medical Consultant

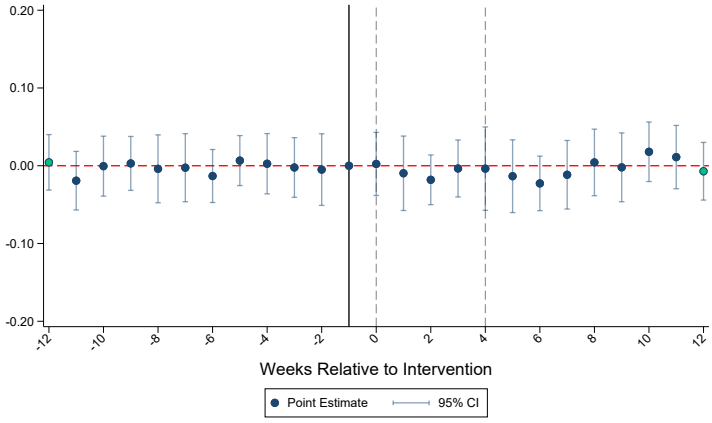


Notes: Each panel corresponds to one of four clinical topics. Bars report the share of topic mentions that the model recommends without any caution (clean recommendations), classified using the stance pipeline described in Appendix B.3. The dark bar is our experiment chatbot; the remaining bars are six comparison models given the same first-turn patient messages under a system prompt framing the model as a direct medical consultant. The dashed horizontal line marks our experiment chatbot's share in each panel. The number of topic mentions (N) appears below each model. Because the follow-up turns of each conversation cannot be reconstructed for the comparison models, our chatbot is restricted here to its first-turn responses, so its shares differ slightly from Table 2, which uses the full analytic sample. Traditional Chinese Medicine mention counts are small for some models, and those shares should be read with caution.

(a) Medication Prescription



(b) TCM



(c) Antibiotics

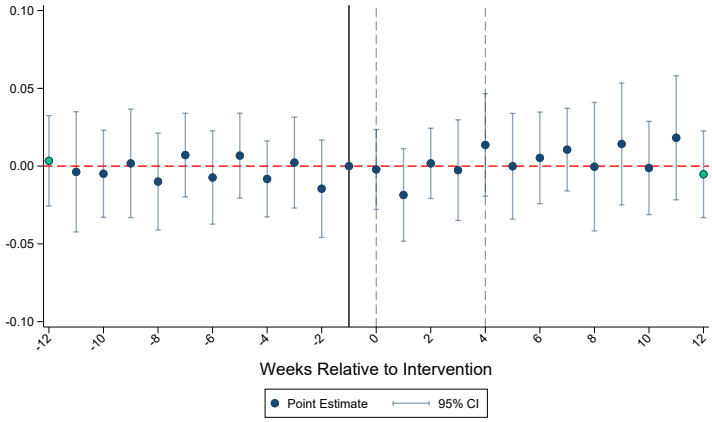
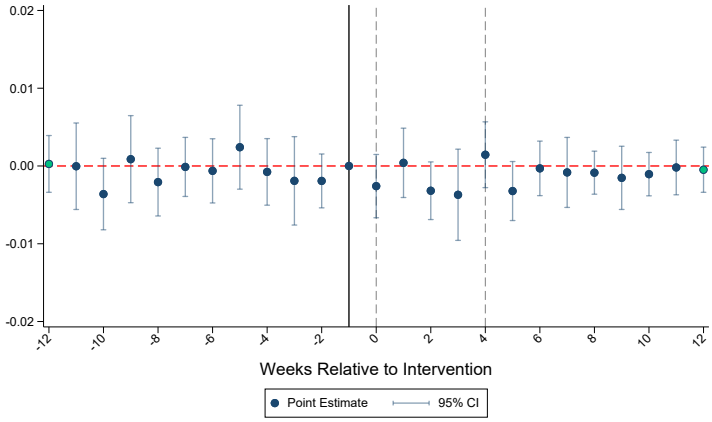
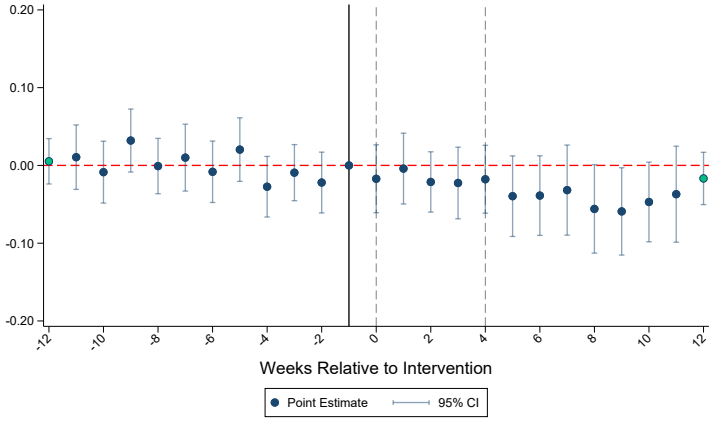


Figure A3. Dynamic Effects of Exposure to Treated Patients on Prescriptions and Clinical Practices, Excluding Treated Patients

(d) Opioids



(e) Diagnostic Tests



(f) Revisit within Two Weeks

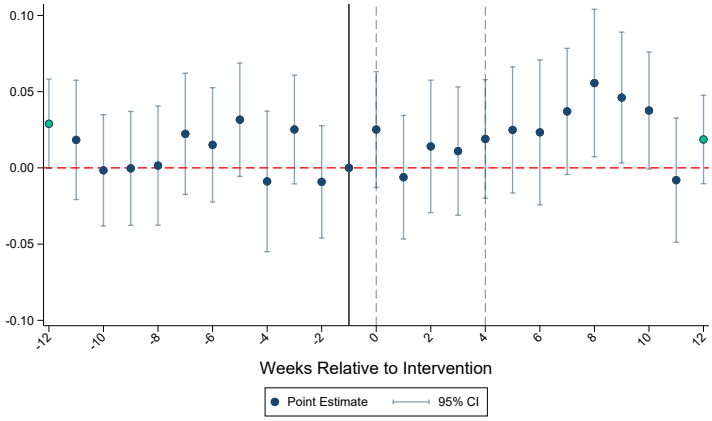
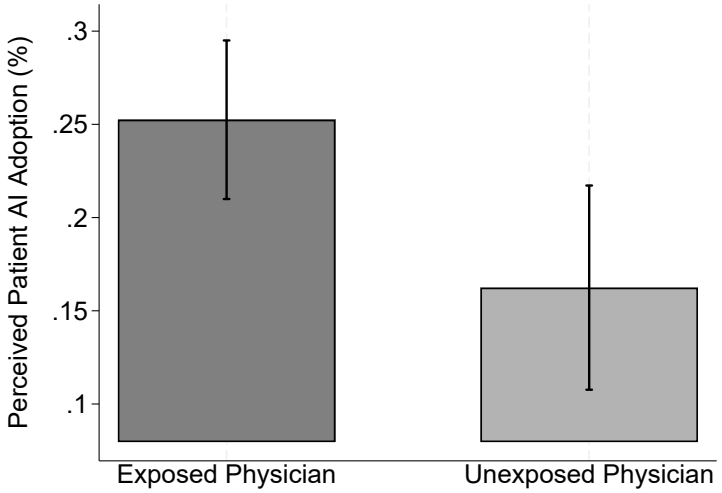


Figure A3. Dynamic Effects of Exposure to Treated Patients on Prescriptions and Clinical Practices, Excluding Treated Patients (Continued)

Notes: This figure presents the point estimates with 95% confidence intervals from equation (2). Data are aggregated to the physician-week level using patient visits from 2025, comparing those from *Control* group and *Unexposed* group, weighted by the number of visits per physician per week. Relative weeks beyond −12 and +12 are grouped at the endpoints. The baseline period is Week −1 (May 26–June 1, 2025). Standard errors are clustered at the physician level.

Figure A4. Physicians' Perceived Patient AI Adoption Rate



Notes: This figure reports physicians' survey responses collected after the conclusion of the experiment, comparing physicians randomized to the *Exposed* and *Unexposed* groups. The outcome variable is physicians' perceived share of patients who used generative AI chatbots prior to consultations during the experimental month, with 95% confidence intervals (dashed lines).

Table A1. Patient–AI Conversation Patterns by Topic and Patient-Message Length

	Mention Rate		AI Stance (% of AI mentions)			
Topic	Patient	AI	Rec Only	Rec+Caution	Caution Only	Neither
Panel A. Least detailed (bottom tercile)						
Diagnostic testing	5.8%	50.0%	93.1%	4.6%	0.0%	2.3%
General medication	7.3%	58.1%	20.5%	43.5%	24.5%	11.5%
Traditional Chinese Medicine	0.6%	9.3%	3.1%	3.1%	84.4%	9.4%
Antibiotics	0.0%	24.1%	3.6%	16.9%	78.3%	1.2%
Panel B. Middle tercile						
Diagnostic testing	4.3%	63.6%	95.3%	1.0%	0.5%	3.1%
General medication	8.9%	65.9%	14.1%	47.7%	31.2%	7.0%
Traditional Chinese Medicine	0.7%	11.9%	2.8%	2.8%	91.7%	2.8%
Antibiotics	0.0%	20.2%	6.6%	11.5%	78.7%	3.3%
Panel C. Most detailed (top tercile)						
Diagnostic testing	14.5%	71.1%	96.8%	1.8%	0.5%	0.9%
General medication	15.8%	63.7%	13.6%	51.3%	24.1%	11.1%
Traditional Chinese Medicine	1.6%	17.4%	5.6%	5.6%	85.2%	3.7%
Antibiotics	0.3%	16.4%	7.8%	5.9%	78.4%	7.8%

Notes: The sample consists of first-turn messages from 956 visits with chatbot access. Terciles are based on the character count of the patient's message. Mention rates are shares of turns in which the topic appears, identified through keyword matching at the sentence level. Patient mentions are based on the patient message; AI mentions are based on the chatbot's reply. The four AI attitude categories are constructed from two independent binary flags assigned by GPT-4o for each turn in which the AI mentions the topic: whether the AI recommends the treatment or test, and whether it raises any caution or advises against its use. Attitude shares are computed as percentages of turns in which the AI mentions the topic and therefore sum to 100% within each row. A single turn can mention multiple topics, in which case it is classified independently for each topic.

Table A2. Effect of Treatment on AI Adoption

	Whether Using AI Consulting Before Visit (1=Yes)	
	(1)	(2)
Treated	0.171***	0.170***
	(0.005)	(0.005)
Patient Characteristics	No	Yes
Observations	11,666	11,666
R-squared	0.093	0.098

Notes: This table reports linear probability estimates of the effect of treatment assignment on AI adoption, using the *Treatment* and *Control* samples. The dependent variable equals one if the patient used the AI chatbot before the visit, and zero otherwise. Column (1) includes only the treatment indicator, while Column (2) additionally controls for patient characteristics, including age, gender, occupation, and an indicator for first visit. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3. Patient Characteristics and AI Usage

	Whether Using AI Consulting Before Visit (1=Yes)			
	(1)	(2)	(3)	(4)
Age	−0.005***	−0.006***	−0.006***	−0.006***
	(0.000)	(0.000)	(0.000)	(0.000)
Male		0.075***	0.055***	0.055***
		(0.009)	(0.012)	(0.012)
Students			−0.001	−0.009
			(0.019)	(0.020)
Farmers			−0.010	−0.008
			(0.013)	(0.013)
Formally Employed			0.067***	0.066***
			(0.013)	(0.014)
Unemployed/Retired			0.024	0.022
			(0.025)	(0.027)
First Visit				0.024**
				(0.010)
Mean	0.171	0.171	0.171	0.171
Observations	5,828	5,828	5,828	5,828

Notes: This table reports estimates from linear probability models examining the relationship between patient characteristics and AI usage. The dependent variable equals one if the patient used AI consulting before the visit, and zero otherwise. The sample is restricted to patients in the treatment group who have access to the chatbot. Columns (1)–(4) sequentially add patient characteristics, including age, gender, student status, occupation, employment status, and first visit status. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4. Intent-to-Treat Effects on the Number of Unique Medications and Diagnostic Tests

	Dependent Variables				
	Number of Medications (1)	Number of TCM Drugs (2)	Number of Antibiotics (3)	Number of Opioids (4)	Number of Diagnostic Tests (5)
Panel A: No Controls					
Chatbot Access	-0.158^{***}	-0.053^{***}	-0.016^{**}	-0.004^{**}	0.071^{***}
	(0.036)	(0.018)	(0.008)	(0.002)	(0.026)
Mean	2.44	0.23	0.14	0.014	0.58
Observations	11,666	11,666	11,666	11,666	11,666
Panel B: Patient Characteristics & Physician FE					
Chatbot Access	-0.160^{***}	-0.054^{***}	-0.015^{*}	-0.004^{*}	0.069^{***}
	(0.036)	(0.019)	(0.008)	(0.002)	(0.026)
Patient Char.	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y
Mean	2.44	0.23	0.14	0.014	0.58
Observations	11,666	11,666	11,666	11,666	11,666

Notes: This table reports intent-to-treat estimates of the effect of chatbot access on the number of unique medications and diagnostic tests prescribed or ordered during a visit. Column (1) reports the number of medications prescribed. Columns (2)–(4) report the numbers of Traditional Chinese Medicine (TCM) drugs, antibiotics, and opioids prescribed, respectively. Column (5) reports the number of diagnostic tests ordered. The sample is restricted to outpatient visits with *Exposed* physicians. Panel A reports estimates without controls. Panel B controls for patient characteristics (age, gender, occupation, and an indicator for first visit) and includes physician fixed effects. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5. Clinical Outcomes by Content of Chatbot Advice

	Dependent Variables			
	General Medication	TCM	Antibiotics	Diagnostic Tests
	(1)	(2)	(3)	(4)
Recommendation Only	0.124***	0.584***	0.118	0.066***
	(0.044)	(0.186)	(0.125)	(0.020)
Caution Only	−0.210***	−0.184***	−0.133***	−0.217***
	(0.041)	(0.007)	(0.006)	(0.020)
Both Recommendation and Caution	−0.083***	−0.206***	−0.142***	0.021
	(0.030)	(0.015)	(0.011)	(0.093)
No Advice	−0.001	−0.017*	−0.012*	0.023***
	(0.009)	(0.009)	(0.006)	(0.008)
Patient Char.	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y
Observations	11,666	11,666	11,666	11,666

Notes: This table reports associations between the content of chatbot advice and clinical decisions in the corresponding category. Each column pairs one topic with its associated decision. General Medication excludes TCM and antibiotics. Patients assigned chatbot access are classified into four mutually exclusive categories for each topic, following the stance flags in Table 2: *Recommendation Only*, *Caution Only*, *Both Recommendation and Caution* and *No Advice*. *No Advice* comprises patients who did not use the chatbot together with users for whom the topic was not raised or was raised without a recommendation or a caution. The omitted category is the *Control* group, which was not assigned chatbot access. All regressions control for patient characteristics (age, gender, occupation, and an indicator for first visit) and include physician fixed effects. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6. Robustness Checks for Intent-to-Treat Effects on Clinical Practice Patterns: First Visit of the Experimental Period Subsample

	Dependent Variables					
	Prescribed medication	TCM	Antibiotics	Opioids	Diagnostic tests	Revisit
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: No Controls						
Treated	−0.036***	−0.018**	−0.011	−0.006*	0.036***	−0.008
	(0.008)	(0.008)	(0.007)	(0.004)	(0.009)	(0.008)
Mean	0.81	0.16	0.12	0.04	0.28	0.14
Observations	8,166	8,166	8,166	8,166	8,166	8,166
Panel B: Patient Characteristics & Physician FE						
Treated	−0.037***	−0.018**	−0.011	−0.006	0.036***	−0.009
	(0.008)	(0.008)	(0.008)	(0.004)	(0.009)	(0.008)
Patient Char.	Y	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y	Y
Mean	0.81	0.16	0.12	0.04	0.28	0.14
Observations	8,166	8,166	8,166	8,166	8,166	8,166

Notes: This table reports within-physician intent-to-treat estimates of the effect of chatbot access on clinical practice patterns. The sample is restricted to patients' first-time outpatient visits with *Exposed* physicians during the experimental period. Panel A reports estimates without controls; Panel B adds patient characteristics (age, gender, and occupation) and physician fixed effects. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7. Robustness Checks for Intent-to-Treat Effects on Clinical Practice Patterns: Clustered Standard Errors

	Dependent Variables					
	Prescribed medication	TCM	Antibiotics	Opioids	Diagnostic tests	Revisit
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: No Controls						
Treated	−0.045***	−0.022***	−0.015***	−0.004	0.027***	−0.011
	(0.007)	(0.005)	(0.006)	(0.003)	(0.008)	(0.008)
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
Panel B: Patient Characteristics & Physician FE						
Treated	−0.046***	−0.022***	−0.015***	−0.004	0.027***	−0.012
	(0.007)	(0.005)	(0.006)	(0.003)	(0.008)	(0.008)
Patient Char.	Y	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666

Notes: This table reports within-physician intent-to-treat estimates of the effect of chatbot access on clinical practice patterns. The sample is restricted to outpatient visits with *Exposed* physicians. Panel A reports estimates without controls; Panel B adds patient characteristics (age, gender, occupation, and an indicator for first visit) and physician fixed effects. Standard errors are clustered at the physician level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8. Robustness Checks for Intent-to-Treat Effects on Healthcare Expenditures: Clustered Standard Errors

	Dependent Variables				
	TCM costs (RMB) (1)	WM costs (RMB) (2)	Diagnostic costs (RMB) (3)	Other costs (RMB) (4)	Total costs (RMB) (5)
Panel A: No Controls					
Treated	−1.686*	−5.276*	6.542	0.174	−0.245
	(1.008)	(3.080)	(7.170)	(3.007)	(8.180)
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666
Panel B: Patient Characteristics & Physician FE					
Treated	−1.762*	−5.021	6.313	−0.002	−0.472
	(0.999)	(3.080)	(7.229)	(3.108)	(8.304)
Patient Char.	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666

Notes: This table reports within-physician intent-to-treat estimates of the effect of chatbot access on expenditures. The sample is restricted to outpatient visits with *Exposed* physicians. Panel A reports estimates without controls; Panel B adds patient characteristics (age, gender, occupation, and an indicator for first visit) and physician fixed effects. Standard errors are clustered at the physician level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9. Robustness Checks for Instrumental Variable Estimates for Clinical Practice Patterns: Clustered Standard Errors

	Dependent Variables					
	Prescribed medication	TCM	Antibiotics	Opioids	Diagnostic tests	Revisit
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: No Controls						
Chatbot Usage	−0.265***	−0.128***	−0.090***	−0.026	0.161***	−0.066
	(0.039)	(0.030)	(0.033)	(0.017)	(0.044)	(0.043)
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
First Stage F-statistics			964.983			
Panel B: Patient Characteristics & Physician FE						
Chatbot Usage	−0.267***	−0.129***	−0.089***	−0.025	0.161***	−0.070
	(0.040)	(0.031)	(0.033)	(0.017)	(0.045)	(0.044)
Patient Char.	Y	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
First Stage F-statistics			947.617			

Notes: This table reports within-physician instrumental variable estimates of the effect of chatbot usage on clinical practice patterns, where chatbot usage is instrumented with random assignment to chatbot access. The sample is restricted to outpatient visits with *Exposed* physicians. Panel A reports estimates without controls; Panel B adds patient characteristics (age, gender, occupation, and an indicator for first visit) and physician fixed effects. The reported first-stage F-statistic is the Kleibergen-Paap Wald F statistic. Standard errors are clustered at the physician level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10. Robustness Checks for Instrumental Variable Estimates for Healthcare Expenditures: Clustered Standard Errors

	Dependent Variables				
	TCM costs (RMB) (1)	WM costs (RMB) (2)	Diagnostic costs (RMB) (3)	Other costs (RMB) (4)	Total costs (RMB) (5)
Panel A: No Controls					
Chatbot Usage	−9.884*	−30.933*	38.359	1.023	−1.436
	(5.911)	(18.072)	(41.813)	(17.560)	(47.759)
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666
First Stage F-statistics			964.983		
Panel B: Patient Characteristics & Physician FE					
Chatbot Usage	−10.334*	−29.444	37.018	−0.009	−2.770
	(5.883)	(18.154)	(42.344)	(18.225)	(48.693)
Patient Char.	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666
First Stage F-statistics			947.617		

Notes: This table reports within-physician instrumental variable estimates of the effect of chatbot usage on healthcare expenditures, where chatbot usage is instrumented with random assignment to chatbot access. The sample is restricted to outpatient visits with *Exposed* physicians. Panel A reports estimates without controls; Panel B adds patient characteristics (age, gender, occupation, and an indicator for first visit) and physician fixed effects. The reported first-stage F-statistic is the Kleibergen-Paap Wald F statistic. Standard errors are clustered at the physician level. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A11. Persistent Effects of Treatment Assignment on Outpatient Utilization in the Post-Experiment Period (October–December 2025)

	Any visit (dummy)		Number of visits	
	(1)	(2)	(3)	(4)
Treated	−0.011	−0.009	−0.022	−0.012
	(0.011)	(0.011)	(0.047)	(0.046)
Patient Char.	N	Y	N	Y
Mean	0.46	0.46	1.24	1.24
Observations	8,166	8,166	8,166	8,166

Notes: This table reports the persistent effects of treatment assignment on patients' use of outpatient care after the experiment ended. The unit of observation is the patient. The sample consists of all patients whose first experimental-period visit was with an *Exposed* physician. We track patients from the experimental period and match their subsequent outpatient visits over October to December 2025 through unique patient identifiers. Because a patient may appear in more than one experimental-period visit with differing assignments, we group each patient by the assignment of their first such visit: *Treatment* if it received chatbot access and *Control* otherwise. The dependent variable in columns (1) and (2) is an indicator equal to one if the patient made at least one outpatient visit during the period; in columns (3) and (4), it is the patient's total visit count over the same period. Columns (2) and (4) include patient characteristics (age, gender, and occupation). Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12. Persistent Effects of Treatment Assignment on Clinical Practice in the Post-Experiment Period (October–December 2025)

	Prescribed medication		Diagnostic tests	
	(1)	(2)	(3)	(4)
Treated	−0.002	−0.003	0.016*	0.016*
	(0.008)	(0.008)	(0.010)	(0.009)
Patient Char.	N	Y	N	Y
Mean	0.83	0.83	0.21	0.21
Observations	10,149	10,149	10,149	10,149

Notes: This table reports persistent effects of treatment assignment on clinical practice at subsequent visits after the experiment ended. The unit of observation is the visit. The sample consists of all outpatient visits over October to December 2025 by patients whose first experimental-period visit was with an *Exposed* physician. Because a patient may appear in more than one experimental-period visit with differing assignments, we group each patient by the assignment of their first such visit, *Treatment* if it received chatbot access and *Control* otherwise. The dependent variables are indicators equal to one if the visit resulted in a prescribed medication, in columns (1) and (2), and in a diagnostic test, in columns (3) and (4). Columns (2) and (4) include patient characteristics (age, gender, and occupation). Standard errors are clustered at the patient level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A13. Effect of Chatbot Access on Patient Perceptions of the Consultation

	Satisfaction (1)	Communication (2)	Compliance (3)
Treated	−0.032	−0.035*	−0.041**
	(0.020)	(0.019)	(0.019)
Patient Char.	Y	Y	Y
Physician FE	Y	Y	Y
Mean	0.86	0.85	0.87
Observations	1,418	1,418	1,418

Notes: This table reports the effect of chatbot access on patient survey responses following the consultation. The dependent variables are binary indicators equal to one if the patient reported (1) being "very satisfied" with the consultation, (2) perceiving physician–patient communication as "very smooth," and (3) intending to "fully comply" with the physician's medical advice. The sample is restricted to survey respondents in the *Treatment* and *Control* groups. All regressions control for patient characteristics (age, gender, occupation, and an indicator for first visit) and include physician fixed effects. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A14. Effect of Patient AI Adoption on Physicians' Perceptions of the Consultation

	Clarity	Understanding	Communication	Compliance
	(1)	(2)	(3)	(4)
Panel A: No Controls				
Share of AI Adopters	0.242	0.228	0.014	−0.215*
	(0.147)	(0.146)	(0.162)	(0.121)
Mean	0.23	0.22	0.35	0.29
Observations	171	171	171	171
Panel B: Physician Characteristics & Department FE				
Share of AI Adopters	0.267*	0.187	0.063	−0.217*
	(0.146)	(0.142)	(0.163)	(0.125)
Mean	0.23	0.22	0.35	0.29
Observations	171	171	171	171

Notes: This table reports the effect of patient AI adoption on physicians' perceptions of the consultation. The dependent variables are binary indicators equal to one if the physician reports that (1) the patient described their symptoms "very clearly"; (2) the patient demonstrated "very good understanding" of their health condition; (3) communication with the patient was "very smooth"; and (4) the patient was "very compliant" with physician instructions. The unit of observation is the physician. *Share of AI Adopters* is the fraction of a physician's patients who used the AI chatbot, equal to zero for *Unexposed* physicians by construction. Panel A includes no controls. Panel B controls for physician characteristics (gender, education, years of experience, and seniority) and includes department fixed effects. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Analyses of Patient–AI Conversation Logs

This appendix provides additional detail on the analyses of the patient–AI conversation logs. The analytic sample contains 1,192 turns from 956 visits. A turn consists of a single patient message and the AI's reply. Most conversations (81.5 percent) consist of a single turn; the remaining 18.5 percent are multi-turn conversations, with two-turn conversations accounting for the majority of these. The analysis proceeds in three steps. We first classify each patient message by its intent, in order to characterize what patients ask the chatbot. We then identify, for each conversation turn, which of four clinical topics the patient and the AI discuss. Finally, for each topic the AI discusses, we classify the AI's stance along two dimensions: whether the AI recommends use of the treatment or test, and whether it raises any caution against use.

B.1 Patient Intent

We classify the intent of each patient message into one of eight categories. Five categories cover the medical content of the consultation: symptom inquiries, in which the patient describes symptoms and asks for a diagnosis or cause; test result inquiries, in which the patient shares a lab result and asks for an interpretation; treatment inquiries, in which the patient asks how to manage a condition or which medication to take; test inquiries, in which the patient asks whether to undergo a specific test; and lifestyle inquiries, in which the patient asks about diet, exercise, or daily self-care. Two categories cover the practicalities of obtaining care: care-seeking, in which the patient asks whether, when, or where to see a doctor, and administrative inquiries about cost, insurance, or scheduling. The descriptive shares reported in Section 4.1 combine these last two into a single "logistical questions" share, since both concern how to obtain care rather than its medical content. A residual "other" category collects greetings and uninformative messages.

Classification is performed by GPT-4o-mini at temperature zero. For each message, the model is given the eight category definitions and returns a primary label, defined as the label that best captures the message's main intent, and zero or more secondary labels. We instruct the model to assign secondary labels only when the message explicitly raises a second distinct topic, not when a second topic is merely implied by the primary.

Appendix Table B15 reports the distribution of primary intent labels across the 1,192 messages

in the analytic sample, together with a representative example of each. Symptom inquiries account for two-thirds of all messages; treatment inquiries are the next largest category, followed by smaller shares of lifestyle, test, care-seeking, result, and administrative inquiries.

Table B15. Primary Intent Distribution and Examples

Intent	N	%	Example patient message (translated)
Symptom inquiry	795	66.7	"Lately I keep feeling tightness in my chest. It hurts a little when I breathe deeply. It's been about a week."
Treatment inquiry	140	11.7	"Do I need to take any medicine? Can I take Suxiao Jiuxin Wan?"
Lifestyle inquiry	65	5.5	"One more question: can I exercise during this period, or will it make symptoms worse?"
Test inquiry	59	4.9	"My foot is still swollen. Should I get an X-ray?"
Care-seeking	51	4.3	"How long do I have to wait before I can get a dental implant?"
Result inquiry	37	3.1	"I'm a 35-year-old woman. A recent physical showed an elevated AST/ALT ratio of 1.64."
Administrative inquiry	28	2.3	"If I get tested at the CDC, do I have to use my real name? Will there be a record?"
Other	17	1.4	"I haven't had that nasal endoscopy."

Notes: Sample of 1,192 patient messages from 956 visits with chatbot access. Each message is assigned exactly one primary intent label. Examples are translated from the original Mandarin.

B.2 Topic Identification

The remaining analysis examines the clinical content of the conversation. We focus on four topics that correspond to the clinical outcomes studied in the main analyses: diagnostic testing, general medication (excluding TCM and antibiotics), Traditional Chinese Medicine, and antibiotics. For each turn, and separately for the patient's message and the AI's reply, we record whether each topic is mentioned.

Topic mentions are identified by keyword matching at the sentence level. We split each message into sentences and check whether any sentence contains a keyword associated with the topic. Sentence-level matching, rather than document-level matching, prevents false positives that arise when distantly related terms co-occur in long structured replies. A turn may match multiple topics,

in which case it enters each topic's analysis independently.

Two topic-specific refinements are worth noting. For diagnostic testing, we distinguish always-specific test names, which match on their own (for example, complete blood count, CT, or endoscopy), from condition terms such as liver function, which match only when paired with a testing-action verb in the same sentence; this prevents general references to a clinical condition from being miscoded as references to a test. For general medication, we exclude any sentence whose match is driven by a TCM keyword, an antibiotic keyword, or a marker for non-pharmacological treatment, so that the residual category captures conventional Western pharmaceuticals without overlapping the TCM and antibiotic categories.

Appendix Table B16 summarizes the keyword scope for each topic.

Table B16. Keyword Lists for Topic Identification

Topic	Keywords
Diagnostic testing	Always-specific test names (e.g., complete blood count, CT, MRI, ultrasound, endoscopy, biopsy) match directly; condition terms (e.g., liver function, lipid panel, thyroid function) match only when paired with a testing-action verb in the same sentence.
General medication	Sentence contains the generic term for medication or a specific drug name (e.g., ibuprofen, acetaminophen, aspirin, metformin, omeprazole, statins, progesterone); excludes sentences also matching TCM, antibiotic, or non-pharmacological keywords.
TCM	Chinese herbal medicine, proprietary Chinese medicines, traditional Chinese medicine practice, herbal decoctions, acupuncture, and tuina massage.
Antibiotics	Generic antibiotic terms and named agents (penicillin, amoxicillin, cephalosporins, azithromycin, erythromycin, levofloxacin, ofloxacin, metronidazole, tinidazole, clarithromycin, roxithromycin, doxycycline, tetracycline). The colloquial Chinese term for "anti-inflammatory medicine," which patients use loosely, is treated as general medication rather than antibiotics.

B.3 AI Stance

For every turn–topic pair in which the AI mentions a topic, we classify the AI's stance along two dimensions: whether the AI recommends the treatment or test, and whether it raises a caution against use. We treat these as independent binary flags. Clinical guidance often combines both: a passage

may recommend acetaminophen for fever above 38.5°C while warning against self-administration of cough suppressants, or recommend physician-supervised use of an antibiotic while discouraging patient-initiated use. Coding the two dimensions separately yields four interpretable cells per topic: recommend without caution, recommend with caution, caution without recommending, and neither.

Classification is performed by GPT-4o at temperature zero. For each turn–topic pair, we extract the topic-matching sentences from the AI reply, concatenate them into a single passage, and submit the passage together with the topic name. The model is instructed to evaluate the two flags only with respect to how the AI frames the specified medical topic.

Three features of the prompt warrant emphasis. First, the prompt specifies that providing dosage, timing, or usage instructions for a treatment counts as recommending it, even when no explicit directive appears, since these instructions carry the same operative meaning as an explicit recommendation. By the same logic, listing tests with diagnostic purposes counts as recommending the tests, and urging the patient to seek care urgently counts as a recommendation rather than a caution. Second, the prompt distinguishes deferring to a physician from recommending physician-supervised use: telling the patient to consult a physician about whether to use a treatment is not coded as a recommendation, while instructing the patient to use a treatment under physician guidance is. Third, the prompt treats negative statements as cautions: stating that a treatment is not needed, not indicated, or not effective for the condition is coded as cautioning, since these statements function as discouragements of use even when the AI does not invoke risk or side effects.

Appendix Table B17 reports representative AI replies illustrating each of the main attitude cells.

B.4 Cross-Model Comparison

To assess whether the directional pattern documented in Section 4.1 is specific to the model our partner deployed, we generate counterfactual responses to the same patient messages from six other large language models and process them through the identical pipeline used for our chatbot. Because the chatbot's later conversational turns depend on patients' reactions to its own earlier replies and cannot be reconstructed for a different model, we restrict the comparison to the first turn of each conversation, that is, the patient's opening message and the model's reply to it. This yields 957 patient messages, to which each comparison model responds once. To match this design, the experiment chatbot is also restricted to its first-turn responses in Figures A1 and A2, so its

Table B17. Illustrative Classification Examples

Topic	Label	Excerpt from AI reply
Diagnostic testing	(yes, no)	If fever persists beyond the third day or worsens, please go to the hospital promptly for a complete blood count and other basic tests to determine whether the infection is viral or bacterial.
General medication	(yes, no)	Oral antihistamines, such as loratadine at a dose of 10 mg once daily, can be used to alleviate itching.
General medication	(yes, yes)	Do not self-administer strong cough suppressants (e.g., those containing codeine). The following is for reference only; specific use must follow physician guidance. For productive cough, ambroxol (Mucosolvan) may be considered as an expectorant. For fever, acetaminophen (Tylenol) is appropriate when temperature is at least 38.5°C.
General medication	(no, yes)	Lifestyle intervention should be the first line of treatment, rather than rushing to medication. If episodes occur more than twice within a year, or if uric acid remains above 540 μmol/L, the patient should seek hospital care to evaluate whether urate-lowering therapy is needed.
TCM	(no, yes)	Self-administration of Chinese herbal medicine or proprietary Chinese medicines to "dissolve nodules" is not recommended; routine follow-up should be the main course of action.
Antibiotics	(no, yes)	Self-use of antibiotics is not recommended. If grinding-related tooth pain has reached the pulp, decompression is required; antibiotics cannot relieve intra-pulp pressure and have no effect on the cavity itself.

Notes: Excerpts are translated from the original Mandarin and lightly condensed. The label column reports (*recommends*, *cautions*).

topic-mention counts and shares differ slightly from Table 2, which uses the full analytic sample.

Models. We use six widely used models from leading developers in the United States and China, listed in Appendix Table B18. Where a model exposes a temperature parameter, we set it to zero for reproducibility; the reasoning models that do not accept the parameter are run at their default setting.

Prompts. To approximate the setting of our intervention, we generate responses under two system prompts. The first frames the model as a pre-visit assistant, matching our deployment; the second frames it as a direct medical consultant. Both instruct the model to respond in Chinese, in free form,

Table B18. Comparison Models

Model	Version	Provider
GPT-4o	`gpt-4o`	OpenAI
GPT-5.5	`gpt-5.5`	OpenAI
DeepSeek-V3	`deepseek-chat-v3-0324`	DeepSeek
Claude Sonnet 4.6	`claude-sonnet-4.6`	Anthropic
Qwen3.5-plus	`qwen3.5-plus-20260420`	Alibaba
Gemini 3.5 Flash	`gemini-3.5-flash`	Google

within approximately 1,000 characters:[37]

> **Pre-visit.** You are a pre-visit medical assistant chatbot at a Chinese hospital outpatient clinic. A patient is about to see a doctor and has sent you a health-related question. Please respond in Chinese, naturally, without following any fixed structure or template, within approximately 1,000 Chinese characters.
>
> **Direct medical consultation.** You are a medical assistant. A patient is consulting you directly for health-related advice. Please respond in Chinese, naturally, without following any fixed structure or template, within approximately 1,000 Chinese characters.

Identical classification pipeline. The comparison models are used only to generate responses. Topic detection and stance classification follow exactly the procedures in Appendices B.2 and B.3: the same sentence-level keyword matching is used to identify topic mentions, and the same GPT-4o classifier, with the same prompt and at temperature zero, assigns the recommendation and caution flags. The classifier receives only the reply text and the topic name, not the identity of the model that produced the reply. The sole difference across the conditions in Figures A1 and A2 is therefore the generating model, which is what allows the figures to attribute differences in stance to the models rather than to the measurement.

Reproducibility. Large language model APIs are not perfectly deterministic, so the same input can yield slightly different responses across runs. To gauge the resulting noise, we randomly select 100 patient messages and pass them through the full pipeline (response generation, topic identification, and stance classification) twice. Comparing the two runs, agreement on the recommendation flag ranges from 88 to 97 percent across models, and agreement on the caution flag from 84 to 98 percent.

[37] Gemini 3.5 Flash occasionally exceeds the requested length limit.

This roughly 10 percent disagreement provides a practical noise floor: differences between models or prompts smaller than this magnitude should not be read as meaningful.