NBER WORKING PAPER SERIES

# TAXPAYER BEHAVIOR IN THE AGE OF AI:
# A FIELD EXPERIMENT ON PROPERTY TAX APPEALS

Justin E. Holz
Ricardo Perez-Truglia
Andrew Simon
Alejandro Zentner

Working Paper 35632
http://www.nber.org/papers/w35632

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2026, Revised August 2026

This Draft: August 18, 2026. This project was reviewed and approved in advance by the Institutional Review Board at the University of Michigan. The field experiment was preregistered in the AEA RCT Registry (AEARCTR-0018430). Upon publication, we will make the replication code and data publicly available through a public repository. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Taxpayer Behavior in the Age of AI: A Field Experiment on Property Tax Appeals
Justin E. Holz, Ricardo Perez-Truglia, Andrew Simon, and Alejandro Zentner
NBER Working Paper No. 35632
August 2026, Revised August 2026
JEL No. C93, H71

**ABSTRACT**

Navigating the tax system often requires information and judgment that taxpayers may lack. Although taxpayers sometimes rely on human experts, AI could provide comparable assistance at substantially lower cost. Yet there is no causal evidence on how taxpayers use AI tax agents or how these tools affect their behavior. We fill this gap with a field experiment in the context of property tax appeals. We recruited a sample of 645 households in Dallas County, Texas. All households received access to a website providing personalized information, filing instructions, and supporting evidence for an appeal. Half were randomly assigned to a version of the website that also included an AI chatbot capable of answering questions and providing guidance tailored to their circumstances. Chatbot take-up was high: 78% initiated a conversation. Access to the chatbot increased the probability of filing an appeal without a human agent by 9.1 pp, from 41.4% to 50.5%. Evidence from clickstream data and conversation transcripts suggests that the chatbot helped households exercise judgment—that is, evaluate and act on the available information. The increase in appeal filing was smaller among less-advantaged households, providing suggestive evidence that the chatbot widened existing disparities. Our results demonstrate the potential of AI tax assistance while highlighting that who benefits depends not only on access, but also on how taxpayers use the technology.

Justin E. Holz
University of Michigan
Ford School of Public Policy
holzj@umich.edu

Ricardo Perez-Truglia
University of California, Los Angeles
and NBER
ricardotruglia@gmail.com

Andrew Simon
University of Virginia
arsimon@virginia.edu

Alejandro Zentner
University of Texas at Dallas
azentner@utdallas.edu

An online appendix is available at
http://www.nber.org/data-appendix/w35632
A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/18430

# 1 Introduction

Interactions between taxpayers and the tax administration help determine how much individuals ultimately pay in taxes. The tax code may specify the deductions, credits, and other benefits to which a taxpayer is entitled, but receiving them often requires the taxpayer to take action—gathering records, learning the relevant rules, completing forms, and sometimes challenging a decision by the tax authority. These steps take time and effort and may also require substantial information and judgment (Benzarti, 2020). As a result, taxpayers may pay more than they legally need to because they fail to claim benefits or pursue remedies available to them (Bhargava and Manoli, 2015). Moreover, because the ability to navigate the tax system varies across individuals, these administrative demands may affect not only average tax burdens but also how those burdens are distributed.

One way taxpayers deal with these challenges is by turning to experts. For example, tax preparers help individuals file their income tax returns and identify deductions and credits, while property tax agents help households challenge their property valuations. These experts do more than complete forms: they translate tax rules into personalized advice and can shape how taxpayers respond to tax incentives and enforcement (Chetty and Saez, 2013; Battaglini et al., 2026). However, access to this assistance can be costly and uneven (Doerner and Ihlanfeldt, 2015). Until recently, these experts have been human.[1] Recent advances in artificial intelligence, however, make it possible for AI systems to provide personalized tax assistance at a much larger scale and at a fraction of the cost.

Purpose-built AI tax agents are still uncommon, but general-purpose chatbots are already being used for tax assistance. For example, in a supplemental survey we conducted with a separate sample of Texas homeowners, 40.0% of those who filed a property tax appeal on their own reported using an AI tool such as ChatGPT for assistance. As these tools become more widely available, policymakers will need evidence not only on who adopts them, but also on how different taxpayers interact with them and how they affect behavior—both on average and across groups. This evidence will be important for anticipating their broader consequences and guiding policy decisions. For example, tax administrations may need to decide whether to offer their own AI agents for free. Policymakers will also need to determine how AI tax agents should be regulated, including who bears responsibility when an agent provides incorrect advice and how taxpayers' privacy should be protected. These decisions

[1] An important intermediate case is conventional tax software, such as TurboTax, which uses pre-specified algorithms to guide taxpayers through common tax situations. For example, an algorithm can calculate the taxpayer's liability using the standard deduction and using itemized deductions, compare the two amounts, and select the option that minimizes the tax due. These tools can be highly effective when the taxpayer's situation falls within the scenarios anticipated by the software's designers, but they are less flexible when taxpayers face unusual circumstances or need assistance with decisions that require judgment.

may ultimately affect who has access to and uses AI tax agents.

We study these questions in the context of property tax appeals in the United States.[2] Households receive a notice from their local appraisal district stating the proposed market value of their home. Households who believe that this value is too high can challenge it—an action Texas law calls a protest and that we refer to as an appeal—either on their own or through a human agent. A successful appeal lowers the ultimately certified value and, consequently, may reduce the property taxes owed, often saving a household hundreds of dollars per year (Nathan et al., 2025). Filing is free and requires completing a short form, either on paper or online. For a household familiar with the process, these mechanical steps can take only a few minutes; for others, the process can still be daunting if they lack relevant knowledge or confidence.

Property tax appeals provide a useful setting for studying AI tax agents for two reasons. First, filing an appeal involves decisions that can benefit from personalized guidance. Households must decide whether an appeal is worthwhile and identify evidence that supports their case. For example, a household whose property has a proposed value of $500,000 may need to determine whether a nearby home that recently sold for less is sufficiently comparable to challenge that value. Because the strength of such evidence depends on the home's characteristics and the information available in each case, property tax appeals offer scope for personalized AI guidance beyond static instructions.

The second reason property tax appeals provide a useful setting is that access to the appeals process, as in many other tax settings, is unequal (Avenancio-León and Howard, 2022; Nathan et al., 2025; Holz et al., 2026). Lower-income homeowners are substantially less likely to file an appeal. Even holding home values constant, Hispanic and Black homeowners appeal at lower rates than White homeowners, while less-educated homeowners appeal at lower rates than more-educated homeowners. These disparities are often attributed, at least in part, to differences in knowledge of the tax system, confidence in navigating the appeals process, and the ability to afford a human agent (Doerner and Ihlanfeldt, 2015). These disparities have led some observers to argue that the "property tax system is rigged against [...] little people" (Lieber, 2020). This setting therefore allows us to study both the average and distributional effects of AI tax assistance. On the one hand, disadvantaged households may have more to gain from an AI agent because it can provide personalized assistance at little or no cost. On the other hand, these households may be less likely to use the agent or to interact with it effectively. Whether AI assistance narrows or widens existing disparities is therefore an empirical question.

[2] Examples of studies on property tax appeals include Avenancio-León and Howard (2022) and Nathan et al. (2025). A broader literature examines property taxation more generally; see, for example, Cabral and Hoxby (2012).

We conducted the field experiment in Dallas County, Texas, because implementing it within a single administrative jurisdiction offered important practical advantages. Dallas County is the second most populous county in Texas, with an estimated population of 2.66 million in 2025 (U.S. Census Bureau, 2026a)—more than the populations of 15 U.S. states (U.S. Census Bureau, 2026b). The county is demographically and politically diverse. Because Texas does not levy an individual state income tax, property taxes are an especially important source of funding for local public services. In 2026, the average Dallas County household was expected to pay approximately $7,900 in property taxes.

We created a website that provided households with personalized property information, step-by-step filing instructions, and supporting evidence for an appeal. In the preregistered experiment, we randomly assigned half of the households to an otherwise identical version of the website that also included an AI chatbot designed to answer questions and provide guidance tailored to their circumstances. Using administrative records, we observed whether each household subsequently filed an appeal. To investigate the mechanisms underlying any effects, we also recorded complete conversation transcripts and tracked households' behavior on the website.

DCAD released its 2026 proposed values on April 14, 2026, and households had until May 15 to file an appeal. Approximately two weeks before the deadline, we mailed postcards to 45,200 households inviting them to use the website. The postcards neither disclosed each household's treatment assignment nor mentioned that the site had two versions. Our analysis sample includes 645 households that visited the website before the deadline.

Engagement with the website was high among households assigned to No-Chatbot, suggesting that households used the available information while considering whether and how to appeal. Chatbot take-up was also high: 78% of households assigned to Chatbot initiated a conversation, often seeking guidance about whether to appeal and how to support their case. Assignment to Chatbot increased the probability of filing a direct appeal from 41.4% in No-Chatbot to 50.5% in Chatbot, a raw difference of 9.1 pp ($p = 0.021$). With a rich set of controls, the estimated effect is similar in magnitude at 10.5 pp ($p = 0.006$). This finding also passes a falsification test based on pretreatment outcomes: when we define the outcome using appeals filed in previous years, we find no effect. Assignment to Chatbot did not crowd out the use of human agents: the probability of filing through a human agent fell by a statistically insignificant 0.2 pp, from 9.6% to 9.4%.

The chatbot effect is not only statistically significant, but also economically meaningful. We benchmark the magnitude of this effect using two prior estimates from the same setting. First, based on estimates from Nathan et al. (2025), the chatbot increases appeal filing as much as an increase of $425 in expected tax savings. Second, in a related field experiment, Nathan et al. (2025) find that mailing households a step-by-step filing guide and a tailored argument based on a nearby comparable home increased the appeal rate by 4.98 pp. Our estimated effect is nearly twice as large. Furthermore, the existing use of general-purpose AI may attenuate our estimated effect.

We explore the causal mechanisms—that is, *why* the chatbot was effective—using clickstream data and conversation transcripts. Assignment to Chatbot increased the average time spent on the website from 12.1 minutes to 17.3 minutes ($p = 0.022$). At the same time, it changed how households used individual website features: it increased the share creating an uploadable evidence document with the PDF Creator but decreased downloads of the pre-generated Comparable Properties Report and use of the step-by-step walkthrough, although the latter two features remained widely used. The conversation transcripts show that households often sought guidance in interpreting the available information, such as determining which comparable homes in the Comparable Properties Report provided the strongest support for their appeal. Taken together, these patterns suggest that the chatbot helped households exercise judgment by navigating the available information, identifying the evidence most relevant to their case, and acting on it.

To investigate whether the chatbot's presentation affected take-up, we conducted a secondary randomization within Chatbot between TaxBot and Marina. TaxBot had a machine-like name and a less personal tone, whereas Marina had a human name and a more personal tone. A more personal presentation might feel more approachable, while a machine-like presentation might make households more comfortable sharing information. Households assigned to TaxBot were a bit more likely to initiate a conversation and file an appeal than those assigned to Marina, but these differences are modest in magnitude and borderline statistically significant or insignificant.

Next, we examine whether the chatbot narrowed or widened existing disparities in the appeal system. The estimated effects on filing rates are smaller among less-educated homeowners, homeowners with lower-valued properties, and homeowners from racial or ethnic minority groups. The subgroup differences are large in magnitude but imprecisely estimated and therefore not always statistically significant. However, when taken together, their consistency across all three dimensions points to a common pattern. Chatbot take-up was also somewhat lower among the same less-advantaged groups, but these differences were considerably smaller than the corresponding differences in filing effects. Thus, take-up cannot fully explain the pattern; much of it appears to arise after take-up, including from differences in how households use the chatbot and translate its guidance into action. Together, these results provide suggestive evidence that universal access to the chatbot may, far from mitigating existing inequities, exacerbate them.

To assess whether these findings were predictable ex ante, we surveyed 87 experts, describing the experiment and asking them to forecast its results. Experts expressed substantial uncertainty and disagreement about the chatbot's effects, which is natural given that this was the first field experiment to provide taxpayers with AI assistance. Experts underestimated chatbot take-up and did not anticipate either the heterogeneity in treatment effects or the relative performance of TaxBot and Marina. However, despite the lack of consensus among experts, their average forecast of the chatbot's effect on filing rates was close to the experimental estimate.

We next discuss how our findings may generalize beyond our study sample. First, our recruitment strategy attracted households who were already considering an appeal. The average household in our sample is not far from the average household in the experimental universe in terms of economic and demographic characteristics, but the sample overrepresents households with a higher propensity to appeal. The effects of the chatbot may therefore differ in the general population. Second, we conducted the experiment in a single county, but property tax appeals operate under the same basic legal framework and follow similar procedures across Texas counties (Nathan et al., 2025). Our findings are also likely to generalize beyond Texas to contexts in which property tax appeals operate similarly, as they do in many other U.S. states and countries (Dobay et al., 2019; World Bank, 2019), although there are exceptions.[3]

Our evidence comes from property tax appeals, an economically important setting in its own right. U.S. property taxes generated an estimated $577 billion in 2019 (Tax Policy Center, 2021), nearly three times as much revenue as the corporate income tax. A broader question is whether our findings extend beyond property tax appeals to other interactions with the tax system. AI tax agents may eventually assist taxpayers in many other settings, perhaps most consequentially with annual income tax filing. Indeed, a survey reported in March 2026 suggests that 26% of people used AI tools, including general-purpose chatbots such as ChatGPT, to obtain help preparing their 2025 tax returns, up from 11% the previous year (Cerullo, 2026). In that setting, taxpayers' circumstances are more varied, and mistakes can carry substantial financial and legal consequences. Those same features, however, create significant technical, legal, and regulatory obstacles to developing and experimentally evaluating a fully functional AI agent.[4] Indeed, at the time of our experiment, major tax-preparation platforms such as TurboTax and H&R Block offered chatbots in some paid tiers,

[3] In some settings, institutional differences may affect households' incentives to appeal and therefore limit extrapolation. In California, for example, Proposition 13 limits increases in assessed values between property sales, often leaving assessments below market values and households with less reason to appeal.

[4] Bock et al. (2025) evaluate frontier models available in 2025 on 51 simplified federal income tax returns. The best-performing model correctly calculated fewer than one-third of the returns. Because AI capabilities are improving rapidly, more recent models may perform better.

but these tools had limited functionality and primarily served as navigation and question-answering assistants. Property tax appeals therefore offer a feasible first setting in which to study AI tax assistance, providing initial evidence that can inform future applications in other tax settings.

Our study relates to and contributes to three strands of literature. First, we provide causal evidence on how access to an AI tax agent affects taxpayer behavior. To the best of our knowledge, ours is the first study to examine this question using a field experiment and administrative data on actual filing decisions.

Second, our study contributes to several bodies of literature on the barriers individuals face when interacting with the government. In addition to the financial burden of taxation, property tax appeals can generate learning and compliance costs because households must decide whether appealing is worthwhile, obtain evidence that fit their case, and prepare an acceptable claim; uncertainty and low confidence may also create psychological costs(Moynihan et al., 2015; Benzarti, 2020; Nathan et al., 2025; Ray et al., 2023). Our findings suggest that AI tax agents can reduce these costs by making filing easier. Complexity and limited information create additional filing frictions that discourage eligible individuals from filing or claiming benefits. Prior work shows that information and assistance can mitigate these frictions in contexts including property tax appeals, tax credits, and food assistance (Nathan et al., 2025; Bhargava and Manoli, 2015; Finkelstein and Notowidigdo, 2019).[5]

We show that AI can reduce filing frictions beyond standard information interventions. Related work also finds that taxpayers may respond ineffectively to tax incentives when they do not understand the relevant rules (Kleven and Kopczuk, 2011; Hoopes et al., 2015; Abeler and Jäger, 2015; Feldman et al., 2016). We show that AI can help taxpayers navigate this complexity through tailored guidance. Finally, we extend studies of human tax professionals and their effects on taxpayer behavior (Chetty and Saez, 2013; Battaglini et al., 2026; DeBacker et al., 2024) by examining AI tax agents.

Third, our study relates to a broader literature on the effects of AI tools. Existing studies show that access to AI tools can increase the productivity of customer-support agents (Brynjolfsson et al., 2025), software developers (Cui et al., 2026), management consultants (Dell'Acqua et al., 2026), and professionals performing occupation-specific writing tasks (Noy and Zhang, 2023).[6] Our setting differs from this existing work. Whereas prior studies focus primarily on productivity, in the tax context the relevant outcomes are hassle costs and tax burdens. Unlike studies finding that AI tools narrow preexisting productivity gaps by

[5] Related work studies pre-populated tax returns, which reduce the information taxpayers must enter and can substantially affect filing costs and tax compliance (Benzarti, 2021; Saez and Seim, 2025).

[6] There are exceptions, however. Otis et al. (2024) find no statistically significant average effect of access to an AI business adviser on revenues or profits among entrepreneurs in Kenya.

disproportionately benefiting less-skilled, less-experienced, or less-educated workers (Noy and Zhang, 2023; Brynjolfsson et al., 2025; Cui et al., 2026; Cruces et al., 2026), our findings suggest that AI assistance may widen existing disparities in property tax appeals.

# 2 Institutional Context

## 2.1 Overview of Property Taxes and Appeals

Local property taxes in Dallas County fund public services provided by the county and by overlapping jurisdictions, including cities, school districts, a community college, a hospital district, and special districts. The Dallas Central Appraisal District (DCAD) is responsible for appraising taxable property, while the relevant jurisdictions set their own tax rates. Each year, DCAD estimates the market value of every property as of January 1. This estimate is known as the property's *proposed value*. Households have the opportunity to appeal that value. After a protest is filed, DCAD reviews the household's requested value and supporting evidence. The district may offer an informal settlement that reduces the proposed value, typically via email. If the household does not accept the offer, or if no offer is made, the case can proceed to a hearing before the Appraisal Review Board. Once protests are resolved, the resulting value becomes the property's certified value, which is used together with applicable tax rates and exemptions to determine the tax liability. As a result, lowering the proposed value through an appeal can generate tax savings in the current and future years. However, in some cases—for example, when the household has a homestead exemption and the homestead cap is binding—the savings may not materialize in the first year.[7]

## 2.2 The Appeal Process and the Role of Evidence

DCAD typically releases proposed values in April, after which households have a 30-day window to file an appeal. Households can check their latest proposed value through DCAD's online portal, and some also receive notices by mail. In 2026, the proposed values were released on April 14, and households had until May 15 to file an appeal. Filing is free. If a household does not attend a scheduled hearing, the protest is dismissed without penalty, and filing ordinarily does not create a risk that DCAD will increase the proposed value.[8] The

[7] The homestead exemption is the most common property tax exemption. Among other provisions, it generally limits annual increases in the appraised value used for taxation to 10%. Other exemptions and caps may also apply; see Nathan et al. (2026) for an overview.

[8] In rare cases, the protest process may reveal inaccurate property characteristics in the public records, prompting DCAD to revise the proposed value upward. For example, DCAD may discover that a home has more bedrooms than its records indicate.

principal costs of filing directly are therefore the opportunity cost of the household's time and the hassle costs of navigating the process.

Households can file on their own, which we refer to as a *direct appeal*, or hire a professional property tax agent to act on their behalf. Agents commonly charge a flat fee, a share of the resulting tax savings, or a combination of the two; some agency agreements continue across multiple years (Nathan et al., 2025). This distinction is important for our analysis. The website and chatbot were designed to assist households who were considering filing on their own; accordingly, direct filing was the primary outcome specified in our RCT preregistration. We also examine appeals filed through human agents because access to the chatbot could crowd out their use.[9]

Households who wish to protest on their own can file online through DCAD's uFile system or submit a paper form by mail; historically, most direct protests have been filed online (Nathan et al., 2025). Appendix A.1 describes the online filing process in greater detail. The form asks households to state the grounds for the protest and allows them to report an "Opinion of Value"—the amount they believe the property was worth as of January 1. Although the mechanical act of submitting the form is straightforward, deciding whether to protest and what claim to make may require information and judgment.

A central feature of this setting is the inherent uncertainty involved in valuing a property that has not recently been sold. DCAD relies on mass-appraisal methods that combine statistical models with information about recent sales and property characteristics. These estimates are necessarily imperfect: two reasonable models may produce different values for the same home, and the models may not capture features or conditions known to the household. Thus, an appeal does not necessarily involve correcting an objectively erroneous proposed value. Instead, the household presents information supporting a lower estimate within a range of plausible values (Nathan et al., 2025).

Households can challenge a proposed value on several grounds. The two most common arguments are that the proposed value exceeds the property's market value and that comparable properties have been appraised at lower values, making the appraisal unequal.[10] Appendix A.1 reproduces the official grounds listed by DCAD and the subsequent steps for uploading evidence and submitting a protest. Supporting evidence can include information about the sale prices or appraised values of comparable properties. The usefulness of a particular comparable depends on factors such as its location, and home characteristics like living

[9] Although this possibility seems unlikely, access to the chatbot could instead increase demand for human agents. For example, the chatbot might persuade a household who otherwise would not have protested that an appeal is worthwhile, after which the household could hire an agent rather than file directly.

[10] Although much less common, the owner can argue that DCAD's records contain errors about the property, such as an incorrect square footage or number of bedrooms.

area, number of bedrooms and bathrooms, age, and condition. Households must therefore decide not only whether they possess evidence, but also which properties are sufficiently comparable, which provide the strongest support, what opinion of value the evidence justifies, and how to present the argument.

### 2.3 Appeal Rates, Tax Savings, and Inequities

In this study, we use publicly available administrative data on property taxes and property tax appeals from the Dallas Central Appraisal District (DCAD) (DCAD, 2026). These data include rich information, such as property addresses and ownership, detailed property characteristics, proposed and certified market values, exemptions, taxable values, tax rates, and historical appeal records.

The average household in Dallas County pays $7,923 in annual property taxes. Property tax appeals can significantly reduce households' tax burden. In 2026, 5.38% of Dallas County households filed an appeal on their own, and an additional 18.35% filed through an agent, for an overall appeal rate of 23.73%. The 2026 appeals are still being processed, but in previous years they often saved households hundreds of dollars per year (Nathan et al., 2025).

The average appeal rate masks substantial heterogeneity. In Dallas County, households from more advantaged groups are substantially more likely to appeal (Nathan et al., 2025), and similar inequities have been documented in other property tax contexts (Avenancio-León and Howard, 2022; Holz et al., 2026). For example, in 2025, the appeal rate was 41.7% among homes in the top quartile of market value (above $513,000), compared with 9.9% among homes in the bottom quartile (below $265,000). Appeal rates are also higher among more-educated and non-minority households.[11] These gaps may reflect differences in expected financial benefits, familiarity with the tax system, confidence in navigating the process, and the ability to afford an agent (Doerner and Ihlanfeldt, 2015).

## 3 Experimental Design and Implementation

### 3.1 The Study Website and Primary Treatment

The experiment was preregistered in the AEA RCT Registry as AEARCTR-0018430. We created a website to help homeowners decide whether to file a property tax appeal and, if so, how to do so. To use the site, households selected their property address. Both versions then

[11] Throughout the paper, we refer to Black and Hispanic households as minority households and White and Asian as non-minority households. This categorization follow prior research showing that, holding home value constant, Black and Hispanic households appeal at lower rates than White and Asian households (Nathan et al., 2025; Holz et al., 2026).

displayed personalized information about the property, including its 2026 proposed value, its value in the previous tax year, and estimated property taxes. Households could also download a "Comparable Properties Report," a PDF containing information about similar properties that could support a market-value or unequal-appraisal claim.[12] The report was intended as a reference tool rather than a ready-to-submit evidence file: it explicitly instructed households not to upload it directly to DCAD and reinforced this warning with a watermark. Finally, the website provided a five-step walkthrough of DCAD's uFile system, with direct links and detailed screenshots explaining how to complete the filing process. Step 3 of the walkthrough also included a simple "PDF Creator": households could paste evidence text into a text box and download the resulting document for upload to uFile. Unlike the pre-generated Comparable Properties Report, which supplied reference information and was not meant to be submitted, the PDF Creator converted household-provided text—potentially drafted with help from the chatbot—into an uploadable evidence document. Figure A.3 reproduces steps 2 through 5; the first step is visible in the full website screenshot in Appendix D. We also include a sample Comparable Properties Report generated by the website in Figure A.5.

The primary randomization assigned households, with equal probability, to one of two conditions: a website without chatbot access—which we refer to as No-Chatbot—and an otherwise identical website with chatbot access—which we refer to as Chatbot. An important feature of the design is that the No-Chatbot group already received substantial assistance: personalized information, supporting evidence, and detailed filing instructions. Households assigned to Chatbot also received access to an AI chatbot powered by Claude, a large language model developed by Anthropic. The chatbot appeared above the personalized property information, so households encountered it before the other website features. They could initiate a conversation by typing a question or selecting one of two quick-start prompts about whether to file an appeal that year or how to file. The chatbot was designed to converse with households and answer questions about their appeals. It received the same property-specific information available elsewhere on the website, as well as information households should already know, such as the property's building characteristics and the household's prior appeal history. As a guardrail, our instructions to the chatbot included a suggested "opinion of value": the lowest value among the property's comparables. This choice anchors the appeal at a defensible value below the current assessment while leaving room to negotiate upward.

The comparison between Chatbot and No-Chatbot is designed to identify the incremental effect of access to the chatbot, above and beyond the assistance provided by the common website. As specified in our preregistration, our central hypothesis is that the chatbot may

[12] The report included up to three recent nearby sales supporting a market-value claim and up to three similar properties in the same DCAD neighborhood that were appraised at lower values, supporting an unequal-appraisal claim. Every property in the experimental universe had at least one admissible comparable.

affect direct appeal filing by helping households determine whether appealing that year is worthwhile and by assisting them in navigating the process. The common website reduced the cost of finding information and evidence, but it did not eliminate decisions requiring judgment—for example, which comparables provided the strongest support, what opinion of value the evidence justified, and how to present the case. The chatbot could help households interpret and act on this information. On the other hand, the chatbot may have little effect if the common website already addresses the most important filing frictions, if households do not trust or engage with it, or if it merely substitutes for other website features without changing their filing decisions.

Consistent with the preregistration, our primary outcome is whether the household filed a direct appeal. We also examine whether households filed through an agent, and we use direct appeals in prior years as placebo outcomes. To study take-up and mechanisms, we observe whether and how households interacted with the chatbot, complete transcripts of their conversations, the time they spent on the website, and the other features they used.

## 3.2 Secondary Treatment: TaxBot versus Marina

Among households assigned to Chatbot, we conducted a second randomization between two versions of the chatbot: TaxBot and Marina. Both versions used the same underlying model, information, and substantive instructions, but differed in their names and personas. TaxBot had a machine-like name and was presented as an automated information system, with instructions to communicate factually, concisely, and procedurally. Marina had a human name and was presented as a friendly and supportive guide, with instructions to communicate warmly, express encouragement, and proactively offer help. Appendix Sections C.2 and C.3 reproduce the complete persona-specific instructions for TaxBot and Marina, respectively. Figure A.4 reproduces screenshots of the two conditions.

The effect of this secondary treatment is theoretically ambiguous. A human name and more personal style could make the chatbot feel more approachable and encourage households to initiate a conversation. Alternatively, a machine-like presentation could appear more appropriate or credible for a technical tax task, or make households more comfortable sharing information with an AI system. We therefore examine whether assignment to TaxBot rather than Marina affected chatbot take-up and, ultimately, direct appeal filing, without imposing a directional prediction.

### 3.3 Recruitment Method via Postcards

We created the postcards after DCAD released its 2026 proposed values and mailed them approximately two weeks before the May 15 filing deadline. The postcard was personalized with the household's name and property address, informed the recipient that the property's 2026 proposed value had been released, and explained that filing an appeal could reduce the household's property-tax burden. It described the website as a free research tool that provided updated property information, helped households decide whether to appeal, and offered step-by-step filing instructions. Recipients could reach the website through either a URL or a QR code. University of Michigan branding and researcher contact information were included to establish the source and legitimacy of the study. Figure A.1 reproduces the front and back of a sample postcard. The postcards were identical across the No-Chatbot and Chatbot groups. In particular, they did not mention the chatbot, disclose the recipient's assignment, or indicate that different versions of the website existed. Thus, households did not learn whether they had been assigned access to the chatbot until they visited the website. Treatment assignment was tied to the property and enforced through the server-side database, preventing households from selecting a different version directly.

We started with a sample of 61,474 owner-occupied, single-family properties in Dallas County whose owners could potentially reduce their property-tax liability by filing an appeal. Among other restrictions, we excluded properties without admissible comparables, properties for which a successful appeal at the suggested value would not reduce the tax bill, and properties whose owners already had professional representation. Appendix A.2 provides additional details about the sample construction. We randomly assigned 16,274 households to a no-mailing group and the remaining 45,200 to receive a postcard inviting them to use the website. Among postcard recipients, 22,600 were assigned to No-Chatbot and 22,600 to Chatbot. Within Chatbot, 11,300 were assigned to TaxBot and 11,300 to Marina. Treatment was assigned by computer at the household level and was not clustered.[13]

[13] We used stratified rerandomization based on prior appeal behavior, suggested savings, property value, voter-file match status, homestead status, proposed-value changes, and the availability of a market-value comparable.

## 3.4 Sample Selection, Descriptive Statistics, and Randomization Balance

Of the 45,200 households to whom we mailed a postcard, 645 visited the website and comprise our main analysis sample.[14] This sample represents 1.43% ($= \frac{645}{45{,}200}$) of the households invited to the website by postcard. Although completing an online survey is not identical to visiting a website for tax assistance, closely related field experiments provide useful benchmarks because they recruited households through mailed invitations containing a URL. Those studies obtained response rates between 1.2% and 3.9% (Nathan et al., 2025; Giaccobasso et al., 2025; Holz et al., 2025, 2026). More broadly, Sinclair et al. (2012) report a response rate of 4.7% for a personally addressed postcard inviting households to complete a web-based survey using a unique code.

Several factors likely contributed to our lower website visitation rate. First, unsolicited correspondence is often discarded or not read carefully (Bottan and Perez-Truglia, 2025; Nathan et al., 2025). Second, the related experiments sent letters enclosed in envelopes, whereas we used postcards to accelerate the printing process; postcards may have been easier to overlook or less likely to inspire trust. Third, logistical constraints required us to mail the postcards only about two weeks before the filing deadline. By then, a substantial share of households who were considering appealing on their own had already done so.[15] Finally, amid recent weakness in the housing market, increases in proposed values were less common than in previous years, likely reducing households' overall interest in appealing. For example, among households in our sampling frame, the share filing an appeal on their own declined from 10.5% in 2025 to 9.4% in 2026.

Because the postcard did not disclose the website version, treatment assignment should not have affected the decision to visit the website. Consistent with this feature of the design, 1.39% of households assigned to No-Chatbot visited the website, compared with 1.46% of those assigned to Chatbot. The difference of 0.08 pp is statistically insignificant ($p = 0.500$). Website visitation also did not differ significantly between TaxBot and Marina—see Appendix A.4 for more details.

Table 1 shows the average characteristics of the analysis sample. Panel (a) reports selected property characteristics and prior appeal behavior from the administrative records, while

[14] The website recorded 1,940 sessions during the study period. In 1,524 of these sessions, the household selected a property address or entered through a direct property link. These sessions corresponded to 710 unique properties. We excluded eight households assigned to the no-mailing group and 57 households who had already filed an appeal before their first website visit. The latter exclusion was specified in the preregistration because the intervention could not affect a filing decision that had already been made.

[15] Among postcard recipients, 9.4% of households ultimately filed an appeal on their own. By comparison, 3.6% had already done so by the time we mailed the postcards. Thus, 38.3% ($= \frac{3.6\%}{9.4\%}$) of those who ultimately filed an appeal on their own had already done so by the time of the mailing.

Panel (b) reports owner demographics for the 79.2% of households successfully matched to a proprietary voter-file dataset.[16] Column (1) describes all website visitors. The average property in the analysis sample had a proposed value of approximately $682,000, an estimated household income of $147,000, and 3.5 bedrooms. About 88% had a homestead exemption, and 61% had multiple owners. Direct appeal rates in the four years before the experiment ranged from 14% to 23%, indicating that many households had prior experience with the process. Among households matched to the L2 voter file, half had a bachelor's degree, 35% belonged to a racial or ethnic minority, 53% were classified as Democrats, and 39% as Republicans.

As discussed in more detail in Appendix A.3, website visitors are roughly similar to postcard recipients on many characteristics. There are some modest differences: visitors own more valuable and larger homes and have higher household incomes and education. The main difference, however, is that website visitors are substantially more likely to have filed an appeal in a prior year. Thus, while the analysis sample remains economically and demographically diverse, it overrepresents households with a higher preexisting propensity to appeal. Related surveys using similar recruitment methods exhibit the same selection pattern (Nathan et al., 2025; Giaccobasso et al., 2025; Holz et al., 2025). For example, in the year before the experiment, 22.6% of website visitors filed an appeal, compared with 10.5% of postcard recipients. The gap in appeal rates is even larger in 2026, as expected: the postcard was designed to attract households considering an appeal that year, and the assistance provided through the website and chatbot was intended to increase filing.

For a randomization balance check, columns (2) through (4) of Table 1 compare households assigned to No-Chatbot and Chatbot. Likewise, columns (5) through (7) compare households assigned to TaxBot and Marina. The differences in characteristics are generally small, although some are statistically significant. Given the large number of tests, we would expect 3.6 of the 36 differences to be statistically significant at the 10% level by chance; we observe somewhat more—six. Out of an abundance of caution, in addition to reporting the raw differences, we present covariate-adjusted estimates using controls selected by post-double-selection lasso (Belloni et al., 2014), hereafter referred to as double lasso. Moreover, we conduct falsification tests in an event-study framework. These checks provide strong reassurance that the treatment effects we document are not spurious.

[16] The L2 voter file provides vendor-modeled measures of household income, education, and race or ethnicity rather than self-reported measures.

### 3.5 Supplemental Surveys

To complement the field experiment, between July and August 2026, we recruited 647 Texas homeowners through Prolific to complete a brief online survey. The survey elicited whether respondents filed a property tax appeal in 2026, whether those who filed on their own used an AI tool such as ChatGPT for assistance, and whether those who did not file had considered appealing or using AI assistance. It also elicited respondents' general use of AI tools and their reasons for not filing or not using AI for an appeal. The median completion time was 4.5 minutes, and 97.5% of respondents passed our AI-agent detection checks. Because these data are self-reported and do not come from the experimental sample, we use them as complementary descriptive evidence on the use of AI tools outside our intervention. Appendix A.7 provides additional details.

We also administered a forecast survey to 87 experts, including academics and county assessors. The survey described the experiment in detail and asked them to forecast several key results. Appendix B describes the expert sample and survey design, while Appendix F reproduces the full survey instrument.

## 4 Results

### 4.1 Chatbot Take-Up

We begin by assessing whether households assigned to Chatbot used the tool and, if so, how intensively they engaged with it. Panel (a) of Figure 1 shows that chatbot take-up was high: about 78% of subjects assigned to the chatbot initiated a conversation. The mean number of messages sent by the subjects was 5.2, with a median of 2—for more details, see panel (a) of Table 2. Thus, for many households, initiating a conversation led to a sustained exchange rather than a single message. These conversations covered a wide range of topics, such as whether it was a good idea to file an appeal that year, what opinion of value to report, and how to write the statement required to accompany the evidence—we return to these data in Section 4.4 below. For reference, we reproduce the full text of two conversations, with sensitive information redacted to protect participants' privacy: Appendix C.4 presents a sample TaxBot conversation, and Appendix C.5 presents a sample Marina conversation.

### 4.2 Effects on Appeal Filing

Even if households had sustained conversations with the chatbot, what matters in the end is whether access to that assistance changed actual tax appeal behavior. The key result is presented in panel (b) of Figure 1: the direct appeal rate increased from 41.4% in No-Chatbot to 50.5% in Chatbot. The raw difference of 9.1 pp is statistically significant (p = 0.021). The figure also reports an estimate that adjusts for pretreatment characteristics selected using double lasso. This adjusted estimate is slightly larger, at 10.5 pp, and statistically significant (p = 0.006).

As a falsification test in the style of an event-study analysis, panel (c) of Figure 1 applies the same analysis to direct appeals filed in prior years. The panel reports the double-lasso-adjusted Chatbot effect separately for each year from 2021 through 2026. Because treatment assignment occurred in 2026, the estimates for 2021–2025 are placebo "effects." All the pretreatment coefficients are close to zero and mostly statistically insignificant.[17] By contrast, the 2026 estimate is positive and statistically significant (p = 0.004). This event-study pattern supports the interpretation that the difference in 2026 appeal rates resulted from access to the chatbot rather than persistent differences in households' preexisting propensity to appeal.

An increase in direct filing need not represent an increase in overall appeal activity if the chatbot merely displaced human agents. Figure 1(d) shows that this did not occur: the probability of filing through an agent fell by a statistically insignificant 0.2 pp, from 9.6% in No-Chatbot to 9.4% in Chatbot (p = 0.935). The adjusted estimate is also small and statistically insignificant: 0.6 pp (p = 0.787). Thus, in our sample, the chatbot increased direct appeals without significantly crowding out appeals filed through human agents.[18]

## 4.3 Economic Magnitude

The effect of the chatbot is not only statistically significant, but also economically significant. We benchmark its magnitude in two ways. First, we translate the effect into a comparable expected financial incentive. Using quasi-experimental variation, Nathan et al. (2025) estimate that every $100 increase in expected tax savings raises the probability of appealing by 2.14 pp. At that rate, expected savings would need to increase by approximately $425 to generate the same 9.1 pp increase in appeal filing as access to the chatbot.

Second, we compare the magnitude of the estimates to the effects of personalized information on how to appeal and provide evidence. In a related field experiment in the same institutional setting, Nathan et al. (2025) find that mailing households a step-by-step filing

[17] The coefficient for 2024 is negative and marginally significant (p = 0.071). Given the large number of falsification tests we conduct, some are likely to appear significant by chance alone.

[18] Crowd-out could be larger under other circumstances. We mailed the postcards approximately two weeks before the filing deadline, by which time households seeking professional assistance may already have hired an agent. Moreover, the mailing sample excluded households with existing professional representation. Providing the chatbot earlier or to households who already used agents could therefore generate greater substitution away from human assistance.

guide and a tailored argument based on a nearby comparable property increased direct appeal filing by 4.98 pp. Our estimate of 9.1 pp is nearly twice as large.[19] This comparison is particularly informative because the No-Chatbot group in our experiment already received detailed filing instructions, personalized information, and a detailed Comparable Properties Report through the website. The chatbot effect therefore represents an increase above and beyond substantial non-AI assistance. Both comparisons come from the same setting—property tax appeals in Dallas County—but should be interpreted with caution because the studies were conducted in different years and their samples, while potentially overlapping, are likely far from identical. With these caveats in mind, the comparisons indicate that the chatbot produced a large behavioral response.

Our primary estimate of 9.1 pp is the intention-to-treat effect of being assigned access to the chatbot, rather than the effect of using it. Because 77.9% of households assigned to the chatbot initiated a conversation, a standard adjustment for incomplete take-up yields a treatment-on-the-treated effect of 11.68 pp ($= \frac{9.1}{0.779}$).[20]

This adjustment accounts for incomplete take-up of our chatbot, but it does not address a separate issue: households assigned to No-Chatbot could still obtain AI assistance outside our website using general-purpose tools such as OpenAI's ChatGPT or Anthropic's Claude. If they did so, the experimental comparison is not between receiving AI assistance and receiving no AI assistance. Instead, it compares access to our chatbot with a control condition in which some households may obtain AI assistance on their own. As a result, even the treatment-on-the-treated estimate may understate the effect of AI assistance relative to a counterfactual with no AI assistance.[21] We conducted the supplemental survey primarily to assess whether such outside AI use was empirically relevant. In a separate sample of Texas households, 40.0% of those who filed an appeal on their own reported using an AI tool for assistance; see Appendix A.7 for more details. The prevalence of AI use among self-filers therefore makes it plausible that the No-Chatbot condition did not correspond to a no-AI counterfactual, which would tend to attenuate the estimated Chatbot–No-Chatbot difference.

[19] We note that the larger effect in our experiment may partially reflect that our sample only includes individuals who visited the website. In a field experiment in Indiana, Holz et al. (2026) find that providing information about how to appeal has a larger effect among website visitors. However, differences in the institutional setting and populations make this comparison more difficult.

[20] This calculation treats initiating a conversation as chatbot take-up and assumes that assignment affects appeal filing only through chatbot use.

[21] Households assigned to Chatbot could also use general-purpose AI tools outside our website, for example because they preferred a familiar tool or had trust or privacy concerns about ours. More generally, outside use of AI makes the experimental contrast less stark than a comparison between AI assistance and no AI assistance.

## 4.4 Mechanisms

Having established that access to the chatbot produced a large and economically meaningful increase in appeal filing, we next examine the mechanisms behind this effect.

We recorded households' behavior on the website, which can shed light on how they engaged with both the website and the chatbot. Figure 2 reports four measures of engagement separately for the No-Chatbot and Chatbot groups: time spent on the site, whether they downloaded the Comparable Properties Report, whether they used the step-by-step walkthrough, and whether they created and downloaded an uploadable evidence document with the PDF Creator.[22] We begin by documenting engagement in the No-Chatbot condition, which provides a baseline for the subsequent comparisons. Engagement was substantial across all four measures: households spent an average of 12.1 minutes on the website, 76.4% downloaded the Comparable Properties Report, 53.2% used the filing walkthrough, and 7.0% created and downloaded an evidence document with the PDF Creator. These patterns suggest that No-Chatbot households did not merely open the website but engaged with the assistance it provided. Moreover, Appendix A.5 shows that households who ultimately filed a direct appeal engaged more extensively than those who did not. Although these comparisons are descriptive, they are consistent with households using the website to decide whether to appeal and, if so, how to do so.

Against this already-high baseline, we assess whether the chatbot changed households' overall engagement with the website. Panel (a) of Figure 2 shows that assignment to Chatbot increased average time on the website from 12.1 to 17.3 minutes. The raw increase of 5.3 minutes is statistically significant ($p = 0.022$), as is the adjusted increase of 5.9 minutes ($p = 0.009$). Thus, the chatbot increased households' overall engagement with the website rather than simply replacing time they would otherwise have spent using the non-Chatbot features. This net effect, however, masks some heterogeneity. Panel (d) shows that the chatbot increased the share creating and downloading an evidence document with the PDF Creator from 7.0% to 18.4%. The raw increase is 11.4 pp ($p < 0.001$), and the adjusted increase is 12.3 pp ($p < 0.001$). On the other hand, the chatbot reduced use of two specific forms of assistance: panel (b) shows that the share downloading the Comparable Properties Report fell from 76.4% to 53.5%, a raw decline of 23.0 pp ($p < 0.001$); the adjusted decline is 21.4 pp ($p < 0.001$). Panel (c) shows that engagement with the step-by-step walkthrough fell from 53.2% to 43.5%, a raw decline of 9.7 pp ($p = 0.014$) and an adjusted decline of 8.8 pp ($p = 0.027$). These features nevertheless remained widely used among households assigned

[22] The website also displayed a table with past and current proposed values and estimated taxes. Because viewing this information did not require households to click on anything, we cannot measure whether they examined it.

to Chatbot. Taken together, these patterns indicate that the chatbot was used mostly in conjunction with the rest of the resources provided on the site.

The clickstream evidence shows that the chatbot changed how households used the website, but it does not reveal what assistance the chatbot supplied. The conversation transcripts provide this complementary evidence. Panel (b) of Table 2 classifies the topics raised during these conversations. The most common question was whether to file an appeal that year, raised by 91.5% of households. Households also asked which comparable properties best supported their cases (50.8%), what opinion of value they should report (44.6%), whether and how to prepare supporting documents or photographs (33.7%), and how to complete the filing process (22.1%). Other conversations addressed the mechanics, costs, and risks of appealing, property characteristics that could support a lower valuation, and what to expect after filing. These exchanges indicate that households used the chatbot not only to retrieve facts, but also to seek advice about decisions requiring judgment.

Panel (c) of Table 2 shows that the chatbot typically responded with concrete guidance. It recommended filing and provided a rationale in 98.1% of conversations, emphasized the filing deadline in 97.7%, and presented or walked households through comparable properties in 85.7%. It recommended an opinion of value in 45.3% of conversations, gave filing or form guidance in 43.8%, and drafted or revised an evidence document in 39.5%. These actions go beyond reproducing static information already available on the website. They suggest that the chatbot helped households exercise judgment by evaluating available information and translating it into action.

## 4.5 TaxBot versus Marina

The preceding results show that households used the chatbot for substantive guidance. We next examine whether take-up and effectiveness depended on how the chatbot was presented. Figure 3 presents the results for the TaxBot and Marina groups, each relative to the common No-Chatbot group. Panel (a) shows that chatbot take-up was higher among households assigned to TaxBot (82.1%) than among those assigned to Marina (74.0%). This raw difference is 8.1 pp and marginally significant ($p = 0.073$). The double-lasso adjusted difference is similar in magnitude at 7.3 pp and is also marginally significant ($p = 0.062$). Thus, the machine-like name and less personal style appear to have encouraged somewhat greater take-up than the human name and more personal style, although the evidence is only suggestive.

Panel (b) of Figure 3 compares the time spent on the website between the TaxBot and Marina conditions and shows that the difference is small and statistically insignificant. Panel (c) compares direct appeal filing rates between the two chatbot conditions. The difference is directionally consistent with the take-up results, with a higher protest rate for TaxBot than for Marina, but the difference is small in magnitude and statistically insignificant. The direct appeal rate was 51.9% among households assigned to TaxBot and 49.1% among those assigned to Marina. The raw difference is 2.8 pp and statistically insignificant (p = 0.619); the adjusted difference is similar in magnitude at 3.5 pp and also statistically insignificant (p = 0.544). For completeness, panel (d) presents the corresponding event-study falsification test. As expected, the Marina–Taxbot differences in the pretreatment years are statistically insignificant. Overall, TaxBot may have been somewhat more effective than Marina at engaging households, but the difference in appeal filing is statistically insignificant. However, this difference is imprecisely estimated, so we cannot rule out modest differences between the two conditions.

## 4.6 Distributional Effects of the Chatbot

Beyond their average effects, policymakers may also be interested in the distributional consequences of AI tax agents. We therefore examine whether the chatbot's effect on appeal filing differed across education, property value, and racial or ethnic minority status. Figure 4 presents the results. Panel (a) begins by comparing take-up rates across each of the three subgroup dimensions. In all three cases, the more advantaged group had somewhat higher take-up, but these differences are always statistically insignificant. Chatbot take-up was higher among more-educated households: 78.4% among those without a bachelor's degree versus 86.2% among those with one (p = 0.106). Take-up was also higher among households with higher-valued properties: 74.5% among those with lower-valued properties versus 79.5% among those with higher-valued properties (p = 0.330). Finally, take-up was higher among non-minority households: 75.4% among minority households versus 79.7% among non-minority households (p = 0.391).

Ultimately, our primary interest is not baseline take-up itself but whether access to the chatbot changed tax behavior. Panels (b)–(d) of Figure 4 display the chatbot's effects on appeal filing across the same three subgroup dimensions.[23] Because several subgroup comparisons rely on smaller samples, we interpret the estimates cautiously and focus on the consistency of the pattern across dimensions rather than any single comparison.

Overall, the chatbot appears to have been less effective among less-advantaged households. More precisely, panel (b) shows that the adjusted effect is 5.8 pp among households without a bachelor's degree (p = 0.345), compared with 25.0 pp among those with at least a bachelor's

[23] For properties with multiple owners, we define demographic characteristics using the first-listed owner in DCAD's ownership records. The BISG imputation uses that owner's surname together with the property's ZIP-code tabulation area, and the match to the L2 voter file selects the registered voter at the property address whose name best matches that owner. Education and minority status therefore refer to the primary owner rather than to the household.

degree ($p < 0.001$). The difference between the two effects is marginally significant ($p = 0.069$). Panel (c) shows an adjusted effect of 5.5 pp among households with lower-valued properties ($p = 0.437$), compared with 13.5 pp among those with higher-valued properties ($p = 0.003$). The raw effects are 3.9 pp and 12.7 pp, respectively. Although the effect for higher-valued properties is more than three times as large, the difference between the two estimates is statistically insignificant ($p = 0.378$). Panel (d) shows the same qualitative pattern by minority status. The adjusted effect is 7.2 pp among racial or ethnic minority households ($p = 0.270$), compared with 14.8 pp among non-minority households ($p = 0.004$); the difference is statistically insignificant ($p = 0.281$).

The subgroup differences are imprecisely estimated, and as a result two of the three are statistically insignificant despite the large magnitude. Nevertheless, the direction and magnitude of the differences are consistent across the three splits. The evidence is therefore suggestive that universal access to AI tax assistance may, far from mitigating existing inequities in the appeal process, exacerbate them. Another important point is that the differences in the chatbot's effects across subgroups are substantially larger for appeal filing (panels (b)–(d) of Figure 4) than for chatbot take-up (panel (a)). This pattern suggests that the distributional effects are driven less by whether households initiated a conversation than by what happened after take-up, including how effectively they translated the chatbot's guidance into action.

Previous studies of non-AI assistance for filing a protest have examined similar dimensions of heterogeneity (Nathan et al., 2025; Holz et al., 2026). Relative to those non-AI interventions, the magnitude and consistency of the widening effects from our AI intervention are noticeably more pronounced.[24]

To assess whether these findings were predictable ex ante, we compare our results with the expert forecasts. Appendix B reports the full results, which we summarize below. The experimental findings were far from obvious ex ante: experts expressed substantial uncertainty, and their forecasts differed considerably. Despite this disagreement, the average forecast of the chatbot's effect on filing rates was close to the experimental estimate. On average, experts substantially underestimated chatbot take-up. There was also no clear consensus about treatment-effect heterogeneity, suggesting that the larger effects among more advantaged households were not clearly anticipated. Finally, experts tended to expect Marina to be more effective than TaxBot, whereas the experimental estimates weakly favor TaxBot.

[24] Nathan et al. (2025) report that the effects of their aid are "if anything, slightly smaller for Hispanic households than for White households," while noting that differences across home-value groups are "mostly small and statistically insignificant." Holz et al. (2026) similarly conclude that their intensive procedural assistance "has almost no additional effect on the racial appeal gap"; their Appendix Figure D4 reports analogous results for the heterogeneity by property value and education.

# 5 Conclusions

This paper provides causal evidence on how taxpayers use AI tax agents and how access to this assistance affects their behavior. We conducted a field experiment with 645 households in Dallas County, Texas. All households received access to a website with personalized information, filing instructions, and supporting evidence for a property tax appeal; half were randomly assigned to receive access to an AI chatbot. Chatbot take-up was high: 78% initiated a conversation. Access to the chatbot increased direct appeal filing by 9.1 pp, from 41.4% to 50.5%, without crowding out human agents. Clickstream data and conversation transcripts suggest that the chatbot was effective because it helped households exercise judgment, evaluating available information and translating it into action. The chatbot had smaller effects on appeal filing among less-advantaged households, providing suggestive evidence that broad access to AI assistance may widen existing disparities. The corresponding differences in chatbot take-up were much smaller, implying that differences in *how* households used the chatbot played a key role.

We conclude by discussing the implications of our findings for policymakers. Our findings point to a broader role for AI in helping individuals navigate the tax system. Tax administrations often try to reduce administrative burdens by providing information, instructions, or online tools. Yet taxpayers may still struggle to determine which information matters and how to act on it. In our experiment, the chatbot affected behavior even though all households already received personalized information, step-by-step instructions, and supporting evidence. This suggests that conversational AI can fill an important gap between making information available and helping taxpayers exercise judgment. Because this assistance can potentially be provided at low marginal cost, tax administrations may want to consider offering AI tax agents for free, both for property tax appeals and for other interactions with the tax system.

Beyond their average benefits, AI tax agents may also have important distributional consequences. In our setting, appeal rates differ substantially across groups, such as between poorer and richer households and between less- and more-educated households. The chatbot had smaller effects on appeal filing among less-advantaged households. Our findings therefore suggest that broad access to AI assistance may widen, rather than narrow, existing disparities. Policymakers should evaluate not only take-up and average effects, but also how different groups use these tools and who ultimately benefits. They may also need to consider whether AI assistance should be paired with more structured guidance or access to human support. These issues will be especially important in more complex settings, such as income tax filing, where inaccurate advice can have substantial consequences and where policies governing accuracy, liability, and privacy may shape both trust and access. As AI tax agents become more common, the key policy question will be not simply whether they work, but for whom and under what conditions.

## References


Abeler, J. and S. Jäger (2015). Complex Tax Incentives. *American Economic Journal: Economic Policy 7*(3), 1–28.

Avenancio-León, C. F. and T. Howard (2022). The Assessment Gap: Racial Inequalities in Property Taxation. *The Quarterly Journal of Economics 137*(3), 1383–1434.

Battaglini, M., L. Guiso, C. Lacava, and E. Patacchini (2026). Tax Professionals and Tax Evasion. *Journal of the European Economic Association, forthcoming.*

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on Treatment Effects After Selection Among High-Dimensional Controls. *Review of Economic Studies 81*(2), 608–650.

Benzarti, Y. (2020). How Taxing Is Tax Filing? Using Revealed Preferences to Estimate Compliance Costs. *American Economic Journal: Economic Policy 12*(4), 38–57.

Benzarti, Y. (2021). Estimating the Costs of Filing Tax Returns and the Potential Savings from Policies Aimed at Reducing These Costs. *Tax Policy and the Economy 35*(1), 55–85.

Bhargava, S. and D. Manoli (2015). Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment. *American Economic Review 105*(11), 3489–3529.

Bock, M. R., K. Molisee, Z. Ozer, and S. Shah (2025). TaxCalcBench: Evaluating Frontier Models on the Tax Calculation Task. arXiv preprint arXiv:2507.16126.

Bottan, N. L. and R. Perez-Truglia (2025). Betting on the House: Subjective Expectations and Market Choices. *American Economic Journal: Applied Economics 17*(1), 459–500.

Brynjolfsson, E., D. Li, and L. Raymond (2025). Generative AI at Work. *The Quarterly Journal of Economics 140*(2), 889–942.

Cabral, M. and C. Hoxby (2012). The Hated Property Tax: Salience, Tax Rates, and Tax Revolts. *NBER Working Paper No. 18514*.

Cerullo, M. (2026). Should You Use AI to File Your Taxes? Experts Warn It Can Lead to Costly Mistakes. *CBS News, March 27 2026*.

Chetty, R. and E. Saez (2013). Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients. *American Economic Journal: Applied Economics 5*(1), 1–31.

Cruces, G., D. Fernández Meijide, S. Galiani, R. H. Gálvez, and M. Lombardi (2026). Does Generative AI Narrow Education-Based Productivity Gaps? Evidence from a Randomized Experiment. *NBER Working Paper No. 34851*.

Cui, K. Z., M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz (2026). The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with

Software Developers. *Management Science*. Articles in Advance, published online February 27, 2026.

DCAD (2026). Appraisal review board data. Dallas Central Appraisal District. https://www.dallascad.org/DataProducts.aspx.

DeBacker, J., B. T. Heim, A. Tran, and A. Yuskavage (2024). Seeking Professional Help: How Paid Preparers Decrease Tax Compliance. *National Tax Journal 77*(2), 229–261.

Dell'Acqua, F., I. McFowland, Edward, E. Mollick, H. Lifshitz, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani (2026). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science 37*(2), 403–423.

Dobay, N., F. Nicely, A. Sanderson, and P. Sanderson (2019). The Best (and Worst) of International Property Tax Administration. Technical report, Council On State Taxation.

Doerner, W. M. and K. R. Ihlanfeldt (2015). The Role of Representative Agents in the Property Tax Appeals Process. *National Tax Journal 68*(1), 59–92.

Feldman, N. E., P. Katuščák, and L. Kawano (2016). Taxpayer Confusion: Evidence from the Child Tax Credit. *American Economic Review 106*(3), 807–835.

Finkelstein, A. and M. J. Notowidigdo (2019). Take-Up and Targeting: Experimental Evidence from SNAP. *The Quarterly Journal of Economics 134*(3), 1505–1556.

Giaccobasso, M., B. Nathan, R. Perez-Truglia, and A. Zentner (2025). Where Do My Tax Dollars Go? Tax Morale Effects of Perceived Government Spending. *American Economic Journal: Applied Economics 17*(4), 223–59.

Holz, J., D. Novgorodsky, and A. Simon (2026). Disparities in tax negotiations: Evidence from field experiments with homeowners and assessors. Working paper.

Holz, J., R. Perez-Truglia, and A. Zentner (2025). Does Tax Avoidance Trickle Down? Evidence from a Field Experiment. *NBER Working Paper No. 34209*.

Hoopes, J. L., D. H. Reck, and J. Slemrod (2015). Taxpayer Search for Information: Implications for Rational Attention. *American Economic Journal: Economic Policy 7*(3), 177–208.

Kleven, H. J. and W. Kopczuk (2011). Transfer Program Complexity and the Take-Up of Social Benefits. *American Economic Journal: Economic Policy 3*(1), 54–90.

Lieber, D. (2020). Why Protesting Your Property Appraisal Is So Hard. *The Dallas Morning News, June 4th 2020*.

Moynihan, D., P. Herd, and H. Harvey (2015). Administrative burden: Learning, psychological, and compliance costs in citizen-state interactions. *Journal of public administration* *research and theory 25*(1), 43–69.

Nathan, B., R. Perez-Truglia, and A. Zentner (2025). My Taxes Are Too Darn High: Why Do Households Protest Their Taxes? *American Economic Journal: Economic Policy 17*(1), 273–310.

Nathan, B., R. Perez-Truglia, and A. Zentner (2026). Paying Your Fair Share: Perceived Fairness and Tax Compliance. *Journal of Accounting and Economics 81*(2), 101838.

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

Otis, N., R. Clarke, S. Delecourt, D. Holtz, and R. Koning (2024). The Uneven Impact of Generative AI on Entrepreneurial Performance. Working paper.

Ray, V., P. Herd, and D. Moynihan (2023). Racialized burdens: Applying racialized organization theory to the administrative state. *Journal of Public Administration Research and Theory 33*(1), 139–152.

Saez, E. and D. Seim (2025). Wealth Tax Enforcement in Sweden: Filing Requirements and Pre-Populated Returns. *Journal of Public Economics 249*, 105440.

Sinclair, M., J. O'Toole, M. Malawaraarachchi, and K. Leder (2012). Comparison of Response Rates and Cost-Effectiveness for a Community-Based Survey: Postal, Internet and Telephone Modes with Generic or Personalised Recruitment Approaches. *BMC Medical Research Methodology 12*(1), 132.

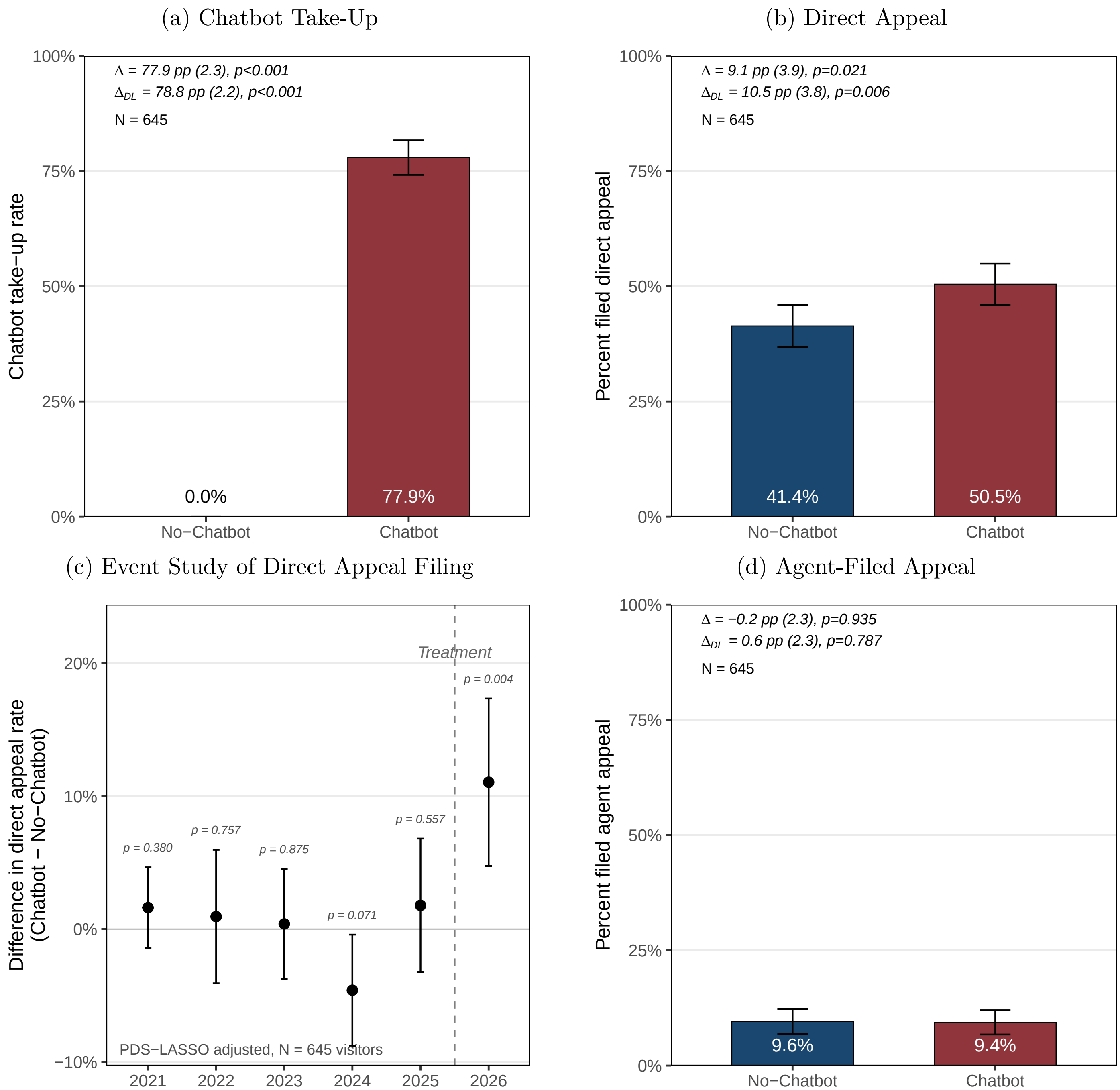
Tax Policy Center (2021). Amount of Revenue by Source. *https://www.taxpolicycenter.org/statistics/property-tax-revenue*.

U.S. Census Bureau (2026a). County Population Totals and Components of Change: 2020–2025. Vintage 2025 population estimates.

U.S. Census Bureau (2026b). State Population Totals and Components of Change: 2020–2025. Vintage 2025 population estimates.

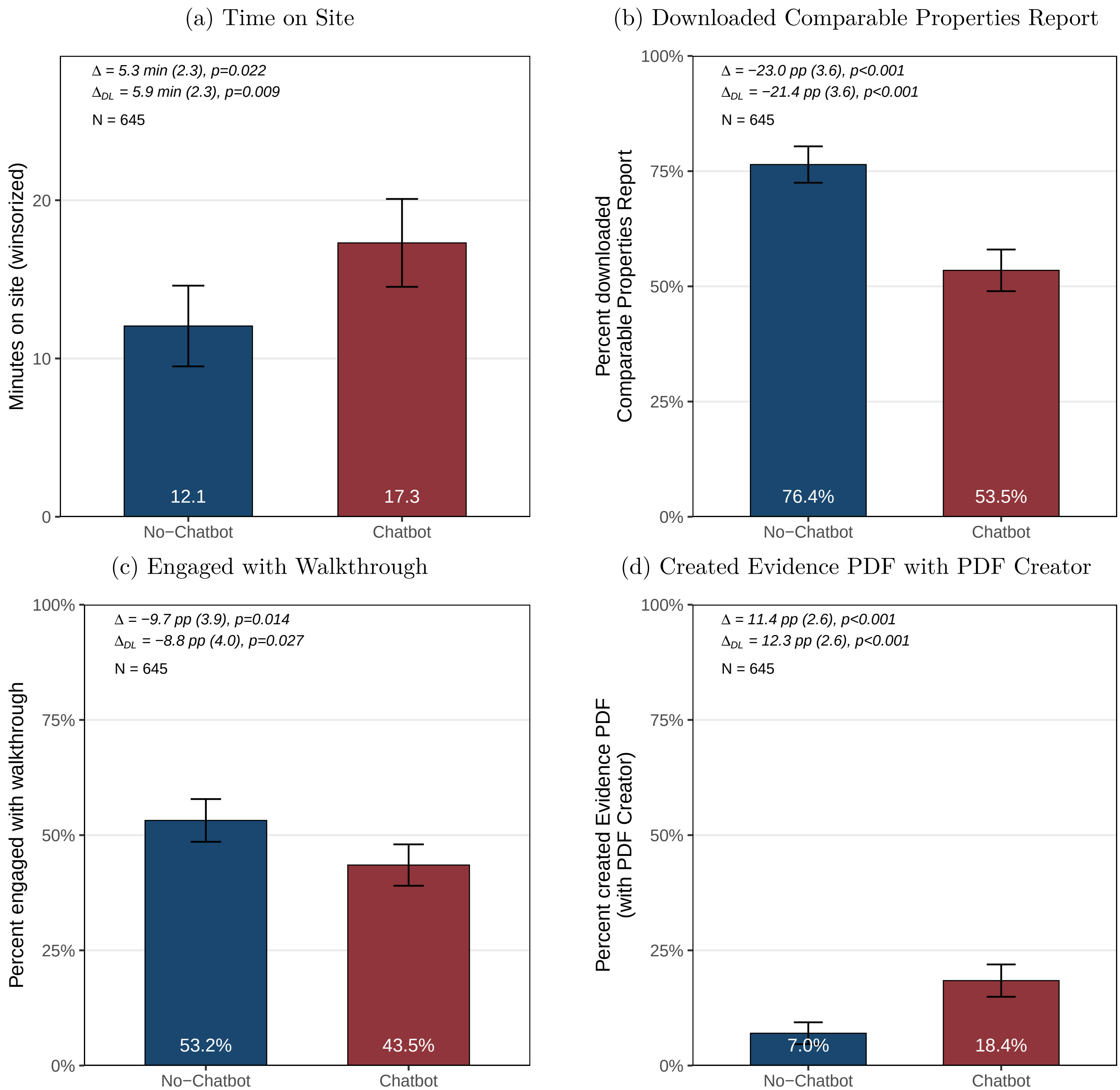
World Bank (2019). The Administrative Review Process for Tax Disputes: Tax Objections and Appeals in Latin America and the Caribbean.

Figure 1: Main Outcomes



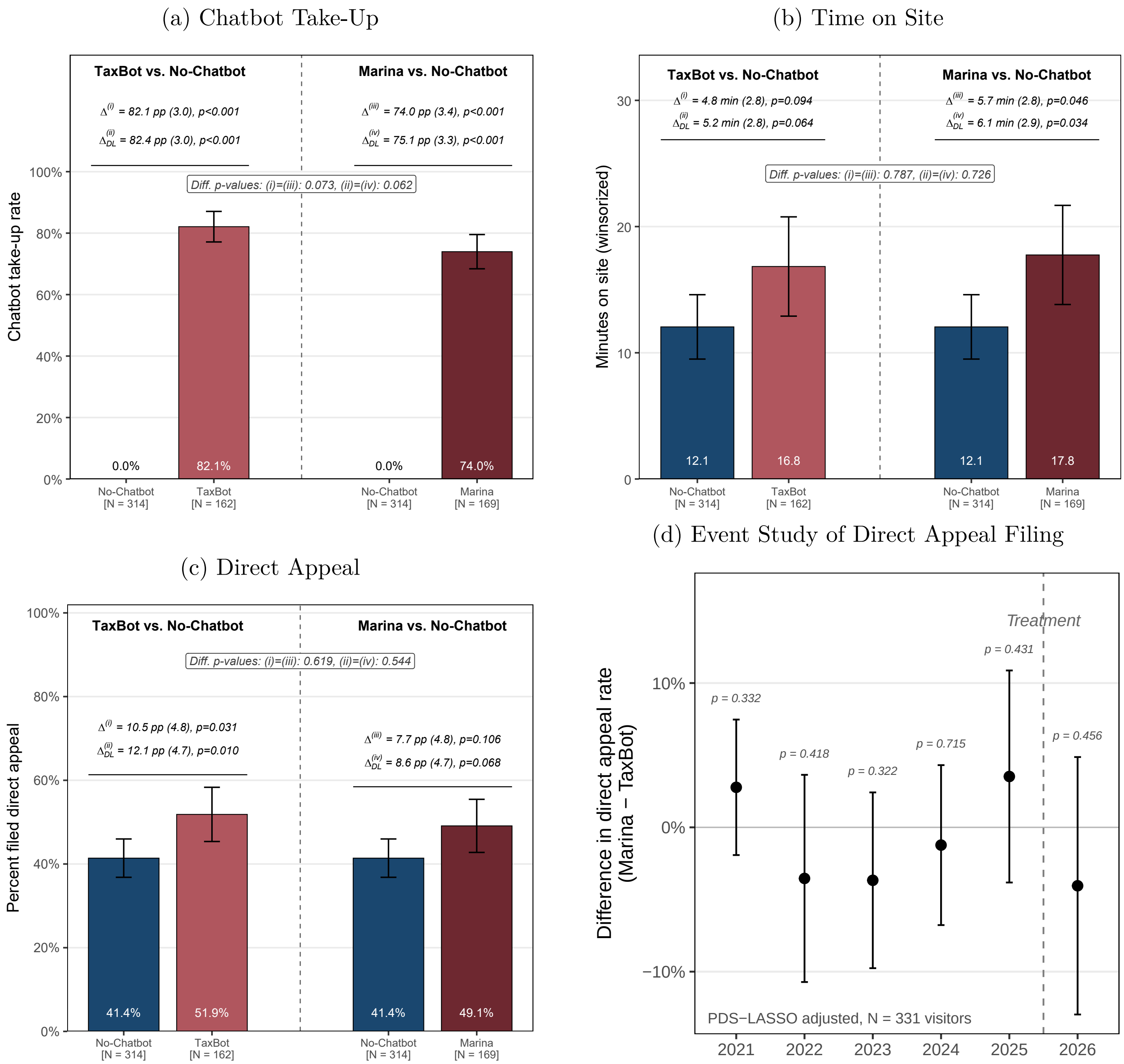

***Notes:*** All panels include website visitors in the analysis sample ($N = 645$). Panels (a), (b), and (d) show the shares who initiated a chatbot conversation, filed a direct (owner-filed, non-agent) appeal, and filed through an agent, respectively. Bars show raw means by arm with 90% confidence intervals and report the Chatbot minus No-Chatbot difference, unadjusted and adjusted with controls selected by double lasso. Panel (c) reports the double-lasso-adjusted effect of Chatbot on direct appeal filing separately by year. Outcomes from 2021–2025 are pretreatment placebos; 2026 is the experimental outcome. Error bars are 90% confidence intervals based on heteroskedasticity-robust standard errors.

Figure 2: Mechanisms: Website Engagement



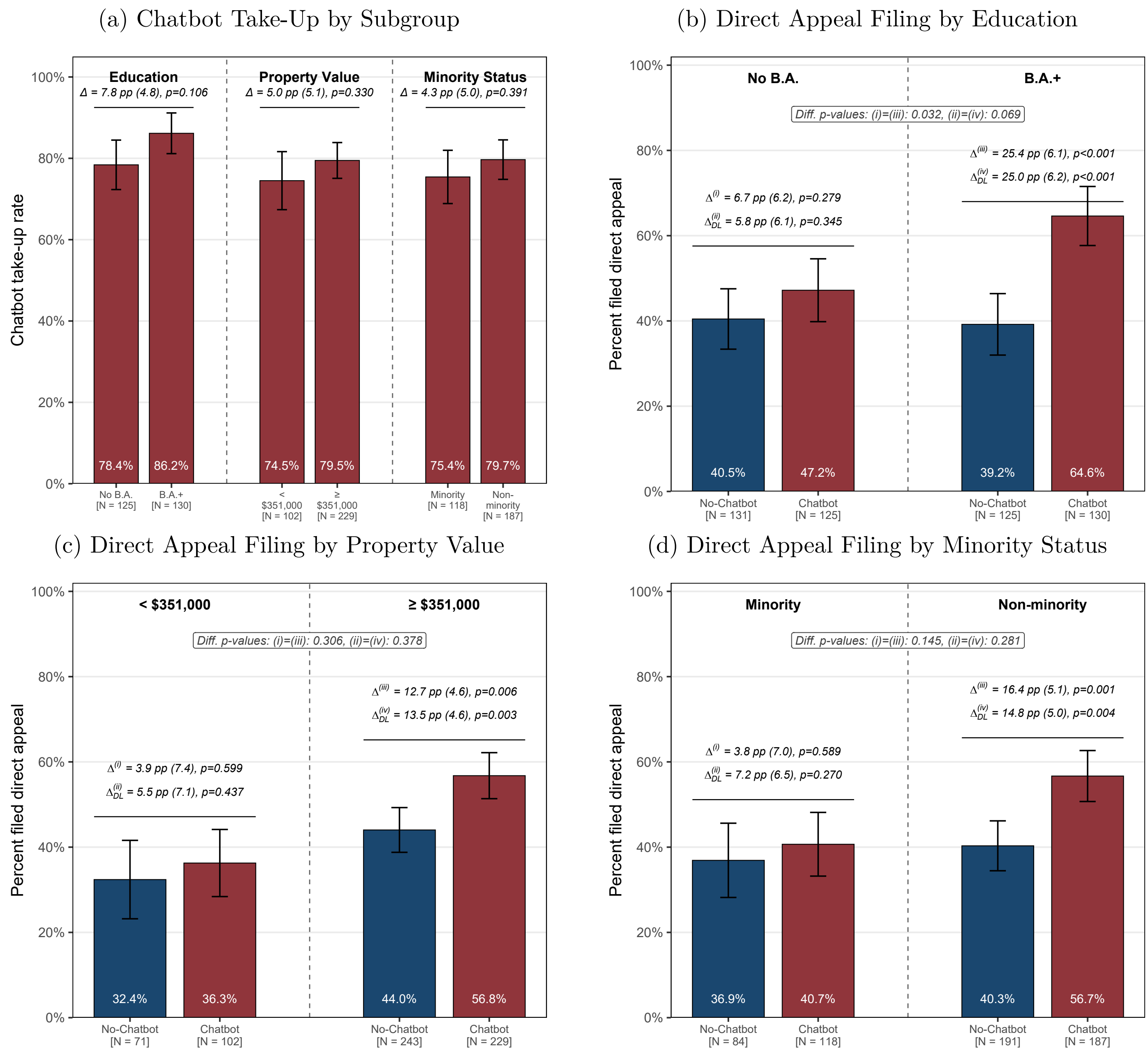

***Notes:*** Includes website visitors in the analysis sample ($N = 645$). Panels follow the order of the website: total minutes on the site (winsorized at the 95th percentile), the share who downloaded the pre-generated Comparable Properties Report, the share who engaged with the step-by-step appeal walkthrough, and the share who created and downloaded an uploadable evidence document with the PDF Creator. Bars show raw means by arm; error bars are 90% confidence intervals. Each panel reports the Chatbot minus No-Chatbot difference (minutes in panel (a), percentage points in panels (b)–(d)), unadjusted and adjusted with controls selected by double lasso, with heteroskedasticity-robust standard errors and p-values from OLS.

Figure 3: Comparison of Chatbot Types: Marina versus TaxBot




***Notes:*** Panels (a) and (c) include all website visitors in the analysis sample ($N = 645$) and show chatbot take-up and direct appeal filing, respectively, for TaxBot and Marina, each compared with the same No-Chatbot benchmark. These panels show raw means with 90% confidence intervals and report unadjusted and double-lasso-adjusted differences. Panel (b) shows minutes on site (winsorized at the 95th percentile) for TaxBot and Marina, each compared with the same No-Chatbot benchmark, with raw and double-lasso-adjusted Marina-minus-Taxbot differences. Panel (d) includes visitors assigned to Chatbot ($N = 331$) and plots double-lasso-adjusted Marina-minus-Taxbot differences in direct appeal filing for 2021–2026. Outcomes from 2021–2025 are pretreatment placebos; 2026 is the experimental outcome. Error bars are 90% confidence intervals based on heteroskedasticity-robust standard errors.

Figure 4: Heterogeneous Chatbot Effects




***Notes:*** Includes website visitors in the analysis sample. Panel (a) includes visitors assigned to Chatbot and reports chatbot take-up rates across the education, property-value, and minority-status subgroups; differences compare the two displayed subgroup means, and brackets report subgroup sample sizes. Panels (b)–(d) show direct appeal filing rates by arm within each subgroup, with 90% confidence intervals. Each subgroup reports the Chatbot minus No-Chatbot difference in percentage points, unadjusted and adjusted with controls selected by double lasso, with heteroskedasticity-robust standard errors and p-values from OLS. Each panel also reports the p-value for equality of the two subgroup effects from a double-lasso-adjusted interaction regression. Education splits on a bachelor's degree from the L2 voter file, conditional on a voter-file match. Property value splits at the median 2026 proposed value among postcard recipients ($350,850). Ethnic/racial minority indicates imputed Black or Hispanic ethnicity via BISG, conditional on BISG coverage, and White and Asian/Pacific Islander owners are non-minority.

Table 1: Descriptive Statistics and Randomization Balance

| | All Visitors (1) | Main Treatment Arm: No-Chatbot (2) | Main Treatment Arm: Chatbot (3) | Main Treatment Arm: $p$-value (2) vs. (3) (4) | Secondary Treatment Arm: TaxBot (5) | Secondary Treatment Arm: Marina (6) | Secondary Treatment Arm: $p$-value (5) vs. (6) (7) |
|---|---|---|---|---|---|---|---|
| *Panel (a): Property characteristics and appeal history* | | | | | | | |
| Proposed value ($1,000s) | 681.5 | 758.3 | 608.6 | 0.002 | 617.9 | 599.7 | 0.753 |
| | (24.3) | (39.2) | (28.7) | | (46.3) | (34.6) | |
| Proposed value change vs. 2025 (%) | 4.8 | 4.0 | 5.6 | 0.381 | 4.0 | 7.2 | 0.323 |
| | (0.9) | (0.7) | (1.6) | | (0.7) | (3.1) | |
| Suggested tax savings ($1,000s) | 1.44 | 1.59 | 1.30 | 0.033 | 1.35 | 1.24 | 0.494 |
| | (0.07) | (0.11) | (0.08) | | (0.13) | (0.09) | |
| Homestead exemption (%) | 88.4 | 90.4 | 86.4 | 0.108 | 87.0 | 85.8 | 0.743 |
| | (1.3) | (1.7) | (1.9) | | (2.6) | (2.7) | |
| Comparable properties available | 3.73 | 3.78 | 3.68 | 0.337 | 3.60 | 3.76 | 0.270 |
| | (0.05) | (0.08) | (0.07) | | (0.10) | (0.10) | |
| Number of bedrooms | 3.49 | 3.51 | 3.47 | 0.430 | 3.42 | 3.51 | 0.234 |
| | (0.03) | (0.04) | (0.04) | | (0.05) | (0.05) | |
| Multiple owners (%) | 60.9 | 62.7 | 59.2 | 0.360 | 53.7 | 64.5 | 0.046 |
| | (1.9) | (2.7) | (2.7) | | (3.9) | (3.7) | |
| 2025 direct appeal rate (%) | 22.6 | 23.6 | 21.8 | 0.583 | 19.8 | 23.7 | 0.389 |
| | (1.6) | (2.4) | (2.3) | | (3.1) | (3.3) | |
| 2024 direct appeal rate (%) | 17.1 | 19.7 | 14.5 | 0.078 | 15.4 | 13.6 | 0.639 |
| | (1.5) | (2.3) | (1.9) | | (2.8) | (2.6) | |
| 2023 direct appeal rate (%) | 14.4 | 15.0 | 13.9 | 0.700 | 15.4 | 12.4 | 0.432 |
| | (1.4) | (2.0) | (1.9) | | (2.8) | (2.5) | |
| 2022 direct appeal rate (%) | 22.8 | 22.9 | 22.7 | 0.935 | 25.3 | 20.1 | 0.262 |
| | (1.7) | (2.4) | (2.3) | | (3.4) | (3.1) | |
| Observations | 645 | 314 | 331 | | 162 | 169 | |
| *Panel (b): Owner demographics* | | | | | | | |
| Matched to L2 voter file (%) | 79.2 | 81.5 | 77.0 | 0.160 | 79.6 | 74.6 | 0.273 |
| | (1.6) | (2.2) | (2.3) | | (3.2) | (3.4) | |
| Household income ($1,000s) | 147.0 | 148.8 | 145.2 | 0.541 | 139.9 | 150.6 | 0.215 |
| | (3.0) | (4.2) | (4.3) | | (5.9) | (6.3) | |
| Female (%) | 33.5 | 33.2 | 33.7 | 0.901 | 40.3 | 27.0 | 0.024 |
| | (2.1) | (2.9) | (3.0) | | (4.3) | (4.0) | |
| Bachelor's degree or more (%) | 49.9 | 48.8 | 51.0 | 0.627 | 48.1 | 54.0 | 0.347 |
| | (2.2) | (3.1) | (3.1) | | (4.4) | (4.5) | |
| Democrat (%) | 53.2 | 53.5 | 52.9 | 0.897 | 55.8 | 50.0 | 0.354 |
| | (2.2) | (3.1) | (3.1) | | (4.4) | (4.5) | |
| Republican (%) | 38.9 | 38.7 | 39.2 | 0.900 | 36.4 | 42.1 | 0.359 |
| | (2.2) | (3.0) | (3.1) | | (4.3) | (4.4) | |
| Ethnic/racial minority (%) | 34.8 | 30.5 | 38.7 | 0.039 | 38.7 | 38.7 | 0.994 |
| | (2.0) | (2.8) | (2.8) | | (4.0) | (3.9) | |
| Observations (matched to L2) | 511 | 256 | 255 | | 129 | 126 | |

***Notes:*** Includes website visitors in the analysis sample. Column (1) pools all visitors; column (2) is the No-Chatbot arm; column (3) pools the two chatbot sub-arms; columns (5) and (6) split the Chatbot arm into TaxBot and Marina. Cells report means with standard errors in parentheses. The $N$ row below Panel (a) reports the number of website visitors in each column, which is the sample for all Panel (a) rows. The $N$ row below Panel (b) reports the number of visitors matched to the L2 voter file, which is the sample for the household income, gender, education, and party rows; the voter-file match and ethnic/racial minority rows use the full visitor sample. Columns (4) and (7) report p-values from two-sample t-tests of the difference between columns (2) and (3), and between columns (5) and (6), respectively. Property characteristics are from DCAD administrative records. The proposed value is the 2026 proposed market value; proposed value change is the percent change from the 2025 to the 2026 proposed value; suggested tax savings is the tax reduction implied by revaluing the property at the lowest-valued comparable property; homestead exemption indicates a 2026 homestead exemption; comparable properties available is the number of comparable properties identified for the homeowner; bedrooms are from the DCAD residential detail file; multiple owners indicates joint owners on the DCAD account. Direct appeal rates in each year are owner-filed (non-agent) appeals from the DCAD ARB archive. Household income, gender, education, and party are from the L2 voter file, conditional on a voter-file match (79% of visitors). Ethnic/racial minority indicates imputed Black or Hispanic ethnicity, based on surname and ZCTA via Bayesian Improved Surname Geocoding (BISG; 92% coverage); White and Asian/Pacific Islander owners are non-minority.

Table 2: Chatbot Conversation Topics and Actions

| | $N$ | Percent |
|---|---|---|
| *Panel A: Conversation Length* | | |
| Number of messages from subject: | | |
| 0 | 0 | 0.0% |
| 1 | 109 | 38.2% |
| 2 | 37 | 13.0% |
| 3 | 16 | 5.6% |
| 4 | 8 | 2.8% |
| 5+ | 115 | 40.4% |
| *Panel B: User topics (% of accounts)* | | |
| **Whether to file protest** | 260 | 91.2% |
| *"Should I file a protest this year?"* | | |
| **Confirmation, feedback, and off-topic** | 137 | 48.1% |
| *"Yes, let's do that. Thank you!"* | | |
| **Comparable property selection** | 135 | 47.4% |
| *"Which comparable homes best support my case?"* | | |
| **Opinion of value questions** | 120 | 42.1% |
| *"What value should I ask for?"* | | |
| **Evidence document and photos** | 90 | 31.6% |
| *"Can you draft the evidence document? Do I need photos?"* | | |
| **Request filing help/instructions** | 60 | 21.1% |
| *"Help me file my protest."* | | |
| **Filing mechanics and form options** | 42 | 14.7% |
| *"Which boxes do I check on the protest form?"* | | |
| **Property features/condition affecting value** | 39 | 13.7% |
| *"My house needs foundation repairs. Does that help my case?"* | | |
| **Value components and corrections** | 23 | 8.1% |
| *"Why did my land value go up so much?"* | | |
| **Process concerns, risks, and cost** | 22 | 7.7% |
| *"Could my value go up if I protest? Does this cost anything?"* | | |
| **Purchase price as basis** | 13 | 4.6% |
| *"I paid much less than the assessed value. Can I use that?"* | | |
| **Post-submission follow-up** | 13 | 4.6% |
| *"I submitted my protest. What happens next?"* | | |
| *Panel C: Chatbot actions (% of accounts)* | | |
| Recommend filing with rationale | 279 | 97.9% |
| Emphasize deadline urgency | 278 | 97.5% |
| Present/walk through comps | 239 | 83.9% |
| Recommend opinion of value | 122 | 42.8% |
| Give filing/form guidance | 118 | 41.4% |
| Draft/revise evidence document | 105 | 36.8% |
| Explain process, risks, and next steps | 85 | 29.8% |
| Correct/clarify input and cost | 43 | 15.1% |
| Acknowledge data limitations / decline out-of-scope | 40 | 14.0% |
| Reframe features/condition as strengthening case | 38 | 13.3% |
| Incorporate user-provided data | 28 | 9.8% |
| Respond to feedback/off-topic | 27 | 9.5% |
| Explain value components and homestead cap | 20 | 7.0% |
| Redirect purchase-price reasoning | 12 | 4.2% |

***Notes:*** Based on households assigned to Chatbot who initiated a conversation ($N$ = 258). Panels (b) and (c) report the number and share of households whose conversations contain each household topic and each chatbot action, classified from full transcripts by a large language model—see Appendix A.6 for details. Topics and actions are not mutually exclusive, so shares do not sum to 100%. The quoted questions in Panel (b) are stylized illustrations of each topic, not verbatim messages.