nature human behaviour

Article

https://doi.org/10.1038/s41562-026-02409-4

The adoption and efficacy of large language models in US consumer financial complaints

Received: 13 March 2025

Accepted: 12 January 2026

Published online: 18 February 2026

Check for updates

Minkyu Shin [1] ✉, Jin Kim [2,3] ✉ & Jiwoong Shin [4] ✉

This research explores the impact of large language models (LLMs) on consumer complaints submitted to the US Consumer Financial Protection Bureau. Analysing 1,134,512 complaints from 2015 to 2024, we document a sharp increase in LLM usage following the release of ChatGPT. An instrumental variable analysis estimates that LLM usage increases the probability of obtaining favourable relief by 6.9 percentage points (95% confidence interval, (4.9, 8.9)). The analysis also reveals evidence of negative selection, where consumers otherwise prone to adverse outcomes are more likely to adopt LLMs. To further substantiate these findings and test the mechanism, we conducted three online controlled experiments (total $N = 1{,}010$ US participants); these demonstrate that LLMs can increase the likelihood of obtaining relief by enhancing the presentation of complaints without altering factual content. These findings suggest that LLMs can act as an equalizer, highlighting the need for policies that expand access to these technologies.

Large language models (LLMs) are rapidly reshaping human communication, offering new tools for enhancing the clarity, precision and persuasiveness of language. This transformation is particularly consequential in high-stakes, asymmetric contexts where individuals must advocate for themselves against large institutions. Whether appealing a health insurance denial, filing immigration paperwork or disputing financial errors, the ability to articulate a compelling narrative is often crucial for securing a favourable outcome. This research investigates the real-world adoption and efficacy of LLMs in one such archetypal setting: consumer complaints in the financial industry. More specifically, we ask two main questions: to what extent are consumers adopting these tools, and does LLM assistance materially improve their outcomes?

The adoption of LLMs is increasing across society[1–3]; however, evidence regarding the extent of consumer adoption remains mixed, partly because most studies rely on survey data, which is subject to substantial measurement error[4]. For example, a survey conducted between November 2023 and January 2024 indicated that ChatGPT is widely used across diverse occupational groups in Denmark, achieving adoption rates as high as 65% among marketing specialists and journalists[5]. In contrast, a survey by the Pew Research Center in July 2023 reported a lower adoption rate of 18% among Americans, which only marginally increased to 23% by February 2024[6,7]. Another study, based on extensive survey data from the USA, shows that 39% of respondents reported using generative artificial intelligence (AI)[8]. Our research documents adoption patterns using a large-scale, non-survey dataset of more than one million consumer complaints spanning nearly a decade, from March 2015 to March 2024. Moreover, by leveraging zip code information in the dataset, we trace the trajectory of these adoption patterns and examine their heterogeneity across various sociodemographic groups.

The efficacy of LLMs in enhancing persuasive outcomes, however, is far from obvious and has yet to be empirically evaluated. While LLMs can enhance the effectiveness of communications by generating content that is coherent, informative and formal[9], they might also diminish effectiveness by producing verbose or redundant content[10] or by failing to capture the subtle nuances of human emotion, which are essential for effective interpersonal communication[11]. Furthermore, the effective use of LLMs requires substantial domain expertise. Crafting precise prompts often demands extensive experience and a thorough understanding of the specific communication context, while assessing the output also necessitates adequate knowledge of the relevant domain. For example, in the finance industry, substantial evidence of a financial illiteracy problem suggests that many consumers may lack the necessary domain knowledge required for effectively leveraging LLMs[12,13]. This deficiency indicates that, even if LLMs can enhance the clarity

[1]City University of Hong Kong, Hong Kong, Hong Kong. [2]Northeastern University, Boston, MA, USA. [3]Chinese University of Hong Kong, Hong Kong, Hong Kong. [4]Yale School of Management, Yale University, New Haven, CT, USA. ✉e-mail: minkshin@cityu.edu.hk; jinkim@cuhk.edu.hk; jiwoong.shin@yale.edu

https://doi.org/10.1038/s41562-026-02409-4

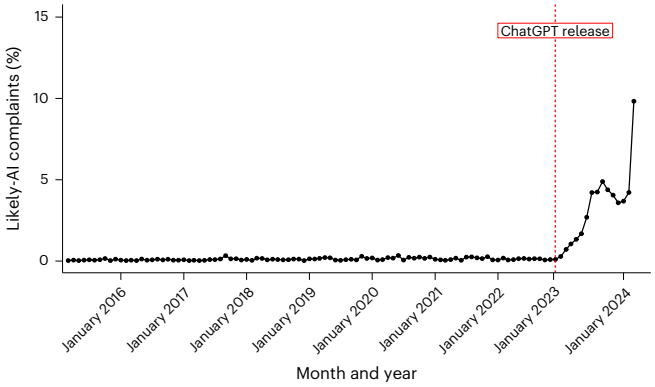

Fig. 1 | Monthly percentage of complaints identified as Likely-AI from a dataset of 1,134,512 CFPB complaints between March 2015 and March 2024. The vertical red dashed line indicates the release of ChatGPT on 30 November 2022. After this release, the proportion of Likely-AI complaints steadily increased, reaching a peak of 9.8% by March 2024.

or linguistic sophistication of consumer messages, a lack of domain knowledge may prevent consumers from using them effectively to persuade financial firms and secure favourable outcomes. Therefore, whether and to what extent LLMs impact the actual outcomes in such scenarios remain open questions.

Our paper, which addresses these two open questions, builds on a nascent but rapidly growing literature on the real-world application of LLMs. Recent research in the human–AI interaction communities, for example, has documented the increasing use of LLMs in various professional domains, such as scientific writing[2], peer review[1,14], workplace communication[15] and online labour markets[16]. However, much of this work focuses on detecting AI usage, measuring its prevalence or examining perceptions of LLM-generated content[17,18]. Our study advances this literature by providing a field-based, causally identified analysis of the actual persuasive impact of LLM-assisted writing. Unlike studies focused on professional-to-professional communication, our setting also involves lay individuals engaging with opaque institutions, where power asymmetries are pronounced and the stakes are personal.

To investigate the adoption and efficacy of using LLMs in consumer communication, we analysed the Consumer Complaint Database from the Consumer Financial Protection Bureau (CFPB) in the USA. This dataset serves as a suitable testbed, providing a large-scale, historical and detailed view of two-sided consumer–firm communication. Our data span nearly a decade from 2015 to 2024, including the pivotal release of ChatGPT on 30 November 2022. This allows us to analyse how both consumer complaints and firm responses changed before and after the introduction of the most popular LLM. The dataset includes over 1.1 million complaints, each containing rich details of the consumer's issue and the firm's response, including the relief decisions made by the firms. We also leverage zip code information for each complaint to merge data from the American Community Survey (ACS) database, enabling us to study the role of sociodemographic factors. The consumer finance domain is particularly apt for this study, as its complex jargon and the prevalence of financial illiteracy create substantial communication barriers for consumers, especially vulnerable ones[12]. This raises the timely question of whether AI can be deployed to empower these consumers by making effective communication more accessible, a topic of growing importance[19].

Our empirical analysis combines observational data, instrumental variable (IV) methods and online controlled experiments to triangulate evidence on the impact of LLMs. Using a validated AI-detection tool, we first documented a sharp increase in LLM usage following the release of ChatGPT, rising from near 0% to 9.8% by March 2024. This early adoption was concentrated in regions with higher proportions of residents who speak a non-English language at home (a proxy for potential language barriers; hereafter, language-barrier proxy). This pattern suggests that language barriers may have contributed to early adoption. We further found that LLM usage increases the likelihood of obtaining relief from financial firms. To address endogeneity concerns, we employed an IV strategy, using zip-code-level Internet access and the language-barrier proxy as instruments for individual adoption. The analysis confirms a positive causal effect of LLM usage, which persists even after controlling for the substantive content of complaints using high-dimensional sentence embeddings. The fact that the effect of LLM use persists even after accounting for what is being said suggests the benefit stems from how it is said—a phenomenon we term the 'presentation effect', by which LLM assistance enhances linguistic cues such as clarity, coherence and professionalism without altering the factual content of complaints. Our IV estimation also provides evidence of negative selection: conditional on observed characteristics, consumers with a lower underlying propensity to obtain relief are more likely to adopt LLM assistance. This suggests that LLMs act as an 'equalizer', providing greater benefits to those who would otherwise be at a disadvantage.

To further substantiate causality and test the mechanism of the presentation effect, we conducted a series of pilot studies and three preregistered online controlled experiments. Three pilot studies developed and validated LLM-edited complaint stimuli, and three preregistered experiments presented hypothetical complaint handlers with complaints that were matched on factual content but varied in clarity, coherence and professionalism. By holding the factual content of complaints constant, these experiments isolate how LLM-driven improvements in clarity, coherence and professionalism, independent of substance, can elicit favourable responses. This process works because LLMs autonomously optimize key linguistic features, enhancing clarity, fluency and formality without requiring domain expertise from the user[20–23]. This finding aligns with established behavioural science theories, such as the elaboration likelihood model[24], suggesting that LLMs enhance the peripheral cues—such as clarity, fluency and professionalism—that evaluators may rely on when forming judgements. This mechanism is also consistent with prior work showing that humans often use such linguistic cues as important heuristics when assessing a text's quality and persuasiveness[25].

Results

Consumer adoption of LLMs for writing complaints in the financial industry

To analyse LLM adoption, we classified complaints as 'Likely-AI' if the detection program assigned them an AI Score greater than or equal to 99%. We selected this conservative threshold on the basis of the bimodal distribution of scores: one peak occurs above 99%, while a much more pronounced peak is observed below 5% (Supplementary Information section 1D). Figure 1 presents the monthly proportion of complaints identified as Likely-AI over time. Prior to the public release of ChatGPT (30 November 2022), the proportion of Likely-AI complaints was minimal, exhibiting only minor fluctuations attributable to false positives. However, in the months immediately following ChatGPT's release, we observe a sharp discontinuity: the proportion of Likely-AI complaints rose steeply from near-zero to 9.8%. We also observe a notable, transient decline in the proportion of Likely-AI complaints around January 2024. While our data do not permit a definitive explanation, this pattern can align with the diffusion-of-innovation theory, potentially reflecting a transitional slowdown between the initial uptake by "early adopters" and subsequent, broader adoption by the "early majority"[26].

Sensitivity analyses using less stringent thresholds (for example, AI Scores ≥80% or ≥70%) revealed an increase in false positives across both pre- and post-release periods. Nevertheless, the surge in the share of Likely-AI complaints remains robust (Supplementary Information section 1D). We therefore retained the conservative threshold of 99%

https://doi.org/10.1038/s41562-026-02409-4

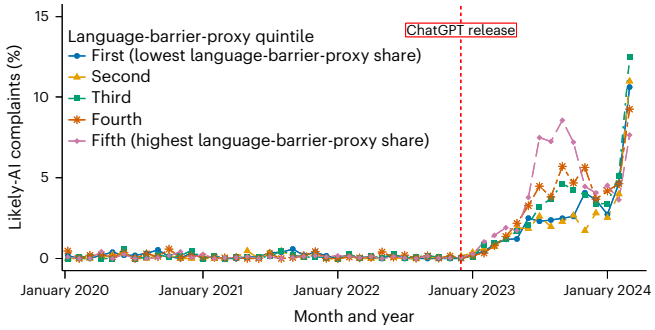

Fig. 2 | Regional variations in LLM adoption from 2020 through 2024, grouped by zip-code quintiles of the language-barrier proxy (zip-code share of residents who speak a non-English language at home; based on the 2017–2021 ACS). Each line/marker corresponds to one quintile from the first (the lowest proxy share) to the fifth (the highest proxy share). The vertical red dashed line indicates ChatGPT's release (30 November 2022). The figure shows that early post-release increases in LLM adoption are larger in higher language-barrier-proxy quintiles.

for the AI Score in our analysis to reliably identify complaints probably generated or assisted by LLMs.

We further observed notable regional heterogeneity in adoption patterns linked to the language-barrier proxy. For this analysis, we divided zip codes of complaints into five quintiles on the basis of the language-barrier proxy (the share of residents who speak a non-English language at home according to ACS data). Following the approach in Fig. 1, we examined temporal trends separately for each quintile. As illustrated in Fig. 2, each line corresponds to one quintile, allowing us to compare the trends across regions with varying levels of the language-barrier proxy. This figure focuses on the period from 2020 to 2024 to highlight adoption patterns; the complete time trend starting from 2015 is available in Supplementary Information section 1E. Adoption was initially concentrated in regions with greater language barriers (the fourth and fifth quintiles), becoming more uniformly distributed across all regions by 2024. This pattern is consistent with language barriers contributing to early adoption, while factors beyond language constraints—such as increased awareness—probably spurred the subsequent expansion. These findings suggest that consumers from diverse sociodemographic backgrounds might have adopted LLMs differently, and this cross-sectional variation in adoption rates served as one of the main IVs in our analysis of LLM efficacy.

The impact of using LLMs on getting relief from financial firms

Having identified the Likely-AI complaints, we tested whether firms would respond more positively to those complaints. We categorized complaints submitted after the release of ChatGPT into two groups: Likely-AI complaints (AI Scores ≥ 99%) and Likely-Human complaints (AI Scores ≤ 5%). We then compared the probability of receiving any form of relief—monetary or non-monetary—from firms between these groups. This comparison revealed a significant difference in success rates (mean for Likely-AI, 49.3%; mean for Likely-Human, 39.9%; difference, 9.4 percentage points; 95% confidence interval (CI), (8.4, 10.4); $\chi^2_{1,244407} = 347.16$; two-sided $P < 0.001$). This positive association remained statistically significant in a logistic regression controlling for time trends and around 200 distinct complaint issue types ($\beta = 0.199$; 95% CI, (0.152, 0.246); $z = 8.292$; two-sided $P < 0.001$); the results from alternative specifications are provided in Supplementary Information section 1F.

While these results suggest a positive association between LLM usage and relief outcomes, addressing potential biases from non-random selection into LLM usage is crucial. Consumers who adopt LLMs may differ systematically from non-adopters in unobserved ways that also influence firm responses. Such characteristics, often correlated with socio-economic status, could bias the estimated effect

Table 1 | First-stage regression results

	Dependent variable: Likely-AI dummy		
	(1)	(2)	(3)
Households with Internet access (share)	0.649		0.884
	(0.173, 1.125)		(0.398, 1.370)
	$z = 2.673$		$z = 3.569$
	$P = 0.008$		$P < 0.001$
Residents speaking non-English at home (share)		0.249	0.297
		(0.126, 0.372)	(0.172, 0.422)
		$z = 3.968$	$z = 4.623$
		$P < 0.001$	$P < 0.001$
Year–month fixed effects	Yes	Yes	Yes
State fixed effects	Yes	Yes	Yes
Issue fixed effects	Yes	Yes	Yes
Product fixed effects	Yes	Yes	Yes
Sociodemographic controls	Yes	Yes	Yes
Observations	244,407	244,407	244,407

This table reports first-stage logistic regression estimates where the dependent variable is a binary indicator for a Likely-AI complaint. The unit of observation is the individual complaint (*i*). The sample period spans December 2022 to March 2024. The values in parentheses represent 95% CIs. All *P* values are two-sided. Sociodemographic controls include zip-code-level median income, educational attainment, employment rate and total household count.

due to endogeneity, implying that the positive association may reflect underlying unobserved consumer attributes rather than the true efficacy of LLMs.

IV estimation. To address concerns regarding endogeneity and non-random selection into LLM usage, we used two socio-economic variables from the ACS as our IVs: Internet access and the language-barrier proxy, both measured at the zip code level. Our strategy of using aggregate-level (zip code) instruments for individual-level adoption decisions follows a well-established practice in applied econometrics[27,28]. These two IVs are based on five-year estimates (2017–2021), preceding the widespread release of LLMs such as ChatGPT, thereby minimizing concerns about contemporaneous correlation with complaint outcomes.

The credibility of our design rests on the exclusion restriction—that these instruments affect relief outcomes only through LLM adoption. To substantiate this restriction[29], we relied on the institutional workflow of complaint handlers, who are instructed to prioritize the complaint narrative over geographic details[30]. Consistent with this focus on content, our model controls for the narrative's substance using high-dimensional embedding vectors as well as issue fixed effects. Although complaint handlers have access to a consumer's billing address, it is implausible that they would access or use zip-code-level statistics—specifically, aggregate Internet access or language-barrier-proxy values—during their review process. Furthermore, to the extent that handlers might infer consumer characteristics from location, our model mitigates this concern by explicitly controlling for the most salient sociodemographic drivers, such as income and education. Finally, our falsification test provides empirical support for this assumption by showing that, prior to the availability of LLMs, there was no evidence of a statistically significant relationship between these IVs and relief outcomes.

Furthermore, each instrument captures a distinct dimension of the adoption decision, leveraging regional variation to predict individual behaviour. Specifically, Internet access proxies for the structural feasibility of adoption: zip codes with Internet connectivity—measured using 2017–2021 ACS data, prior to the widespread release of LLMs—provide core infrastructure that facilitates LLM

https://doi.org/10.1038/s41562-026-02409-4

Table 2 | Second-stage regression results

	Dependent variable: relief dummy			
	(1)	(2)	(3)	(4)
Likely-AI dummy	0.338	0.338	0.339	0.069
	(0.228, 0.448)	(0.222, 0.454)	(0.225, 0.453)	(0.049, 0.089)
	$z = 6.023$	$z = 5.711$	$z = 5.828$	$z = 6.762$
	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$
First-stage residual ($\hat{r}_i$)	−0.031	−0.031	−0.031	−0.005
	(−0.049, −0.013)	(−0.047, −0.015)	(−0.049, −0.013)	(−0.007, −0.003)
	$z = -3.376$	$z = -3.798$	$z = -3.376$	$z = -4.900$
	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$
First-stage IV	Internet	Language-barrier proxy	Both	Both
Second-stage model	Logit	Logit	Logit	Linear
Year–month fixed effects	Yes	Yes	Yes	Yes
State fixed effects	Yes	Yes	Yes	Yes
Issue fixed effects	Yes	Yes	Yes	Yes
Product fixed effects	Yes	Yes	Yes	Yes
Sociodemographic controls	Yes	Yes	Yes	Yes
Sentence-embedding controls	Yes	Yes	Yes	Yes
Observations	244,407	244,407	244,407	244,407

This table reports second-stage estimates where the dependent variable is a binary indicator for whether a company provided relief. Columns 1–3 report logistic regression estimates, and Column 4 reports linear probability model estimates. The unit of observation is the individual complaint (i). The sample period spans December 2022 to March 2024. The first-stage residual ($\hat{r}_i$) is included as a control function to address the endogeneity of LLM usage. The values in parentheses represent 95% CIs. All P values are two-sided.

adoption, whereas limited connectivity increases the friction and cost of access. In contrast, the language-barrier proxy captures the regional intensity of demand: in areas with greater language barriers, the marginal benefit of linguistic assistance is higher across many writing tasks, thereby increasing the likelihood that individuals leverage LLMs when drafting complaints. Consistent with this interpretation, both instruments significantly predict LLM adoption in the first stage (Table 1). Together, these instruments isolate variation in LLM usage driven by both supply-side opportunity and demand-side need. Further details on the estimation procedure, specifically the two-stage residual inclusion (2SRI) model, are provided in the Methods.

Table 1 reports estimates from the first-stage logistic regressions testing whether our IVs predict individual LLM adoption. Column 1 indicates that in zip codes with a higher proportion of households with Internet access, consumers are more likely to submit complaints classified as Likely-AI ($\beta = 0.649$; 95% CI, (0.173, 1.125); $z = 2.673$; two-sided $P = 0.008$). Similarly, Column 2 shows that in zip codes with a greater language-barrier proxy, consumers are more likely to submit Likely-AI complaints ($\beta = 0.249$; 95% CI, (0.126, 0.372); $z = 3.968$; two-sided $P < 0.001$). Column 3 corroborates these findings, confirming that both instruments remain statistically significant predictors when included simultaneously (Internet access: $\beta = 0.884$; 95% CI, (0.398, 1.370); $z = 3.569$; two-sided $P < 0.001$; language-barrier proxy: $\beta = 0.297$; 95% CI, (0.172, 0.422); $z = 4.623$; two-sided $P < 0.001$). These first-stage results indicate that our IVs are empirically relevant predictors of LLM usage even after controlling for a comprehensive set of fixed effects and covariates. We included the residual from this first-stage regression in the second-stage equation as a control function to account for potential non-random selection into LLM usage.

Table 2 presents the main results from the second-stage regressions, showing that consumers who used LLMs to compose their complaints are more likely to receive relief. This relationship holds after accounting for a wide range of potential confounders, including time trends, state-level heterogeneity, zip-code-level sociodemographics, specific complaint issues and product categories. Columns 1–3 report estimates from the 2SRI models using alternative first-stage instrument sets (Internet access only, language-barrier proxy only and both), while Column 4 presents estimates from a linear probability model using both instruments.

Across all specifications (Columns 1–4), the primary coefficient of interest—the Likely-AI dummy—is positive and statistically significant (Column 1: $\beta = 0.338$; 95% CI, (0.228, 0.448); $z = 6.023$; two-sided $P < 0.001$; Column 2: $\beta = 0.338$; 95% CI, (0.222, 0.454); $z = 5.711$; two-sided $P < 0.001$; Column 3: $\beta = 0.339$; 95% CI, (0.225, 0.453); $z = 5.828$; two-sided $P < 0.001$; Column 4: $\beta = 0.069$; 95% CI, (0.049, 0.089); $z = 6.762$; two-sided $P < 0.001$). This indicates that, under our identification strategy, LLM usage is associated with a higher probability of securing relief even after accounting for selection on observables and unobservables. Specifically, the linear probability model (Column 4) estimates an increase of approximately 6.9 percentage points ($\beta = 0.069$; 95% CI, (0.049, 0.089); $z = 6.762$; two-sided $P < 0.001$) in the probability of success for Likely-AI complaints. This estimate remains stable when we include year–month, state, product and issue fixed effects, demographic covariates, and high-dimensional sentence embeddings that control for the semantic content of complaints.

The coefficient on the first-stage residual term—which serves as the control function in the 2SRI framework—provides information about the selection mechanism. This coefficient captures the correlation between unobserved determinants of LLM adoption and unobserved determinants of complaint success. Its statistical significance is consistent with the presence of endogeneity. Across all models, we observe a negative and statistically significant coefficient on the residual term (for example, Column 3: $\beta = -0.031$; 95% CI, (−0.049, −0.013); $z = -3.376$; two-sided $P < 0.001$), indicating negative selection: consumers whose unobserved characteristics make them more likely to adopt LLMs are, on average, those who would be less likely to obtain relief in the absence of LLM assistance.

This pattern of negative selection is consistent with the view that LLMs do not merely amplify the advantages of already well-positioned consumers; rather, they may act as an equalizer. Conditional on observed characteristics, LLM adoption is higher among consumers

https://doi.org/10.1038/s41562-026-02409-4

Table 3 | Falsification test using pre-ChatGPT data

	Dependent variable: relief dummy		
	(1)	(2)	(3)
Households with Internet access (share)	−0.068		−0.064
	(−0.256, 0.120)		(−0.254, 0.126)
	z = −0.709		z = −0.660
	P = 0.478		P = 0.509
Residents speaking non-English at home (share)		0.010	0.007
		(−0.043, 0.063)	(−0.048, 0.062)
		z = 0.370	z = 0.249
		P = 0.712	P = 0.803
Year–month fixed effects	Yes	Yes	Yes
State fixed effects	Yes	Yes	Yes
Issue fixed effects	Yes	Yes	Yes
Product fixed effects	Yes	Yes	Yes
Sociodemographic controls	Yes	Yes	Yes
Sentence-embedding controls	Yes	Yes	Yes
Observations	624,588	624,588	624,588

This table presents reduced-form regression estimates using data from the pre-ChatGPT period (January 2017 to October 2022). The dependent variable is a binary indicator for company-provided relief. The unit of observation is the individual complaint (i). The values in parentheses represent 95% CIs. All P values are two-sided. The lack of statistical significance across all specifications supports the exclusion restriction by indicating that the instruments did not directly influence relief decisions prior to the availability of LLM tools.

who—absent LLM assistance—would be less likely to obtain relief. By enhancing the presentation of complaints, LLMs may therefore help level the playing field for consumers who otherwise face disadvantages. Under negative selection, non-IV correlational estimates with extensive observable control variables may be biased downward relative to the causal effect, because adopters tend to have a lower baseline probability of success without LLM assistance.

Falsification test. If our IVs affected relief outcomes through channels other than LLM adoption, the exclusion restriction would be violated, and our causal interpretation of the IV estimates would be undermined. To probe this possibility, we conducted a falsification test using pre-ChatGPT data—that is, a period when LLMs could not be used to draft complaints and thus could not influence relief decisions, while any alternative pathways from the instruments to relief decisions would still be present. Under the exclusion restriction, the instruments should not exhibit systematic predictive power for relief decisions in this pre-LLM period. Falsification tests of this type are standard in the IV literature for assessing instrument validity by examining whether alternative pathways or pre-existing relationships violate the exclusion restriction[31–33].

Table 3 presents the results of this falsification test, with each column corresponding to a specification analogous to those in Tables 1 and 2. Across all specifications, we did not observe statistically significant associations between the IVs and relief decisions. For example, the coefficient on Internet access in Column 1 is −0.068 (95% CI, (−0.256, 0.120); z = −0.709; two-sided P = 0.478), and the coefficient on the language-barrier proxy in Column 2 is 0.010 (95% CI, (−0.043, 0.063); z = 0.370; two-sided P = 0.712). When both instruments are included (Column 3), the estimates remain statistically non-significant (Internet access: −0.064; 95% CI, (−0.254, 0.126); z = −0.660; two-sided P = 0.509; language-barrier proxy: 0.007; 95% CI, (−0.048, 0.062); z = 0.249; two-sided P = 0.803). These null results indicate that, in the absence of LLMs, there is no evidence that the instruments statistically predict relief outcomes. While this test cannot rule out all possible direct effects, the failure to detect a systematic relationship across multiple specifications is consistent with the view that, in the post-ChatGPT era, our estimated effects primarily operate through LLM usage rather than through pre-existing channels.

While our IV analysis, together with this falsification test, provides plausibly causal evidence that LLM usage improves complaint outcomes, an important limitation remains. The observational data cannot fully capture the extensive margin—the possibility that LLMs draw in a new set of consumers who would not otherwise submit a complaint. If such consumers may differ systematically in the nature of their cases, such compositional changes could complicate causal inference. To address this concern and to isolate the effect of LLM assistance itself—holding constant both the complainer and the substance of the grievance—we conducted online controlled experiments.

These experiments were designed not only to corroborate causality but also to elucidate the underlying mechanism. Our observational models already controlled for the core content of each complaint using issue-specific fixed effects and high-dimensional sentence-embedding vectors. The fact that the effect of LLM use persists even after accounting for what is being said suggests the benefit may stem from how it is said. This pattern aligns with the peripheral route to persuasion, in which linguistic attributes—such as clarity, fluency and professionalism—influence outcomes independently of a message's core arguments. Accordingly, our experiments directly tested this 'presentation effect' by experimentally manipulating the stylistic attributes of complaints while holding their factual content constant.

Controlled experiments testing the presentation-enhancing effect of LLM editing

By directly manipulating the presentation of complaint narratives in a structured environment, controlled experiments can further validate our findings from IV regressions and explore a possible mechanism underlying firms' differential responses to LLM-assisted complaints. To investigate this relationship in greater depth, we conducted three pilot studies and three preregistered experiments with participants recruited from Prolific (see Supplementary Information section 5 for details on participant recruitment, compensation and demographics by study). We analysed data from all participants who completed the study and did not exclude anyone. Across all pilot studies and experiments, the participants were compensated at an average rate of US$10.19 per hour (see Supplementary Table 9 for study-specific compensation rates).

The three experiments were similar in design and collectively replicated and extended the core findings. We present the experiments in an order that reflects conceptual progression and ease of understanding rather than the chronological order in which they were conducted. Across all the experiments, we employed a within-participant design, which increases precision by controlling for stable rater-specific tendencies and enhances ecological validity by reflecting how complaint handlers evaluate multiple narratives in practice. Further details on the experimental design are provided in Supplementary Information section 4. The preregistration, materials, data and analysis code for each pilot study and experiment are hosted on the project's Open Science Framework page.

A summary of three pilot studies. Three pilot studies were conducted to achieve distinct objectives. Pilot Study 1 validated our experimental design by testing whether participants' judgements reproduced decision-making patterns observed in the CFPB data. Pilot Study 2 identified the presentation features participants most wanted to improve using LLMs. Pilot Study 3 evaluated the extent to which LLMs actually enhanced complaint presentation. Further details on each pilot study are available in Supplementary Information section 2.

The primary aim of Pilot Study 1 (N = 216; 48.1% men, 50.0% women, 1.9% other; mean age, 36 years) was to validate that our experimental set-up approximated the decision-making of financial firms. We selected 108 complaints from the CFPB data, all written before the release of LLMs (that is, before 2022), half of which had resulted in monetary relief and half of which had not. The participants each evaluated a random set of five complaints and rated the likelihood of offering monetary compensation on a 1–7 scale. Their judgements closely tracked firms' actual decisions: complaints that had received monetary relief were significantly more likely to be offered compensation than those that had not (mean, 4.75 versus 4.20); a linear fixed-effects model with participant fixed effects confirmed this difference (β = 0.53; 95% CI, (0.33, 0.73); s.e. = 0.10; t_{863} = 5.13; P < 0.001 (two-sided)). This result supports the validity of our experimental design and suggests that participant evaluations align with the judgements of decision makers at financial firms.

Pilot Study 2 examined how people might use LLMs to improve the presentation of their complaints. We recruited 212 participants from Prolific (N = 212; 49.1% men, 49.1% women, 1.9% other; mean age, 40 years) and asked them to imagine writing complaints "to obtain monetary relief from a financial institution". Each participant rated five complaints sampled from the CFPB data on their intention to improve ten linguistic features. Participants expressed the strongest intention to improve clarity (mean, 3.78), coherence (mean, 3.71) and professionalism (mean, 3.71) and weaker intention to improve features such as assertiveness (mean, 3.05) and politeness (mean, 2.86). See Supplementary Information section 2B and Supplementary Table 6. Notably, these participant-selected dimensions map onto distinct scholarly traditions: clarity aligns with readability research and Grice's maxim of manner (sentence level)[34], coherence with discourse cohesion[35] and rhetorical structure theory (paragraph/document level), and professionalism with register theory and politeness research (pragmatic–situational level). Guided by these empirical and theoretical insights, we developed stimuli for Experiments 1 and 2. Specifically, we selected 20 complaints from the CFPB data and used ChatGPT to edit each to enhance one of the three top-rated features: clarity, coherence or professionalism. This yielded 80 total complaints: 20 unedited (Control) and 60 edited (20 Clear, 20 Coherent and 20 Professional).

Pilot Study 3 tested whether the 60 edited complaints were enhanced in presentation compared with the 20 unedited ones. Participants (N = 490; 49.4% men, 50.0% women, 0.6% other; mean age, 39 years) were randomly assigned to evaluate complaints on clarity, coherence or professionalism. (A note on sample size: although we preregistered and asked Prolific for a sample size of 480 participants, Prolific provided us with 490 completed responses, all of which were included in the analyses.) The results showed that the edited complaints scored higher on all three dimensions (Fig. 3 and Supplementary Fig. 9). Specifically, three separate linear fixed-effects models controlling for participant fixed effects and complaint (content) fixed effects confirmed these enhancements in presentation; see Supplementary Table 13 (Models 1–3). Thus, even though each participant assessed only one linguistic feature, we found that LLM editing improved overall presentation quality across multiple linguistic features simultaneously[20,21]. This validation supports the use of these edited complaints in our main experiments to examine how stylistic improvements can affect relief decisions.

Experiment 1. In Experiment 1 (N = 301; 48.8% men, 49.8% women, 1.3% other; mean age, 42 years), each participant evaluated five distinct complaints randomly selected from the pool of 80 (no complaint was repeated per participant; Supplementary Information section 3A). For each complaint, they rated the likelihood of offering monetary compensation on a 1–7 scale, which served as the dependent measure.

The results of Experiment 1 align with our findings from the CFPB data. We primarily focus on the preregistered linear fixed-effects model (Model 2 in Table 4), which controls for the effects of participants and complaint contents. This model showed that the complaints edited with ChatGPT (Coherent, Clear and Professional complaints) were more likely to receive hypothetical monetary compensation than the unedited (Control) complaints (β = 0.306; 95% CI, (0.135, 0.476); s.e. = 0.087; t_{1184} = 3.52; P < 0.001 (two-sided); see Model 2 in Table 4). Additional linear models controlling for different sets of fixed effects (Columns 1, 3 and 4 in Table 4) consistently indicate that editing complaints with ChatGPT significantly increased the likelihood of hypothetical compensation. Moreover, this pattern held regardless of which linguistic feature the edits focused on (all Holm-adjusted two-sided P < 0.018; see Model 4 in Table 4). These results corroborate the findings from the section 'The impact of using LLMs on getting relief from financial firms', confirming that editing complaints with an LLM significantly increases the likelihood of obtaining relief.

Experiment 2. In Experiment 2 (N = 303; 49.8% men, 49.8% women, 0.3% other; mean age, 44 years), we replicated Experiment 1 with a more ecologically valid sample. Specifically, from Prolific, we recruited individuals who had prior work experience in the finance sector and had completed a college degree or higher. Aside from this difference in the participant pool, all other aspects of the study, including stimuli and design, were identical to Experiment 1.

We analysed the data following the preregistration (https://aspredicted.org/mc5k-n58z.pdf). As predicted, a linear fixed-effects model revealed that complaints edited with ChatGPT (to improve clarity, coherence or professionalism) were significantly more likely to receive hypothetical monetary compensation than unedited complaints (β = 0.32; 95% CI, (0.16, 0.49); s.e. = 0.086; t_{1192} = 3.78; two-sided P < 0.001; see Model 2 in Supplementary Table 7). The results were robust across alternative model specifications (Supplementary Table 7) and support our hypothesis in a sample that more closely resembles actual complaint handlers (as observed in the CFPB data) than the general public.

Experiment 3. Experiment 3 (N = 406; 49.0% men, 50.5% women, 0.5% other; mean age, 37 years) tested whether the positive effect of LLM editing on compensation likelihood observed in Experiment 1 would replicate when the length of edited and unedited complaints was controlled. We constructed a new set of 100 complaints: 50 unedited complaints drawn from Pilot Study 1 and 50 corresponding LLM-edited versions created by asking ChatGPT to improve clarity. Unlike Experiments 1 and 2, the edited and unedited complaints were matched in length, with no significant difference in mean character count.

https://doi.org/10.1038/s41562-026-02409-4

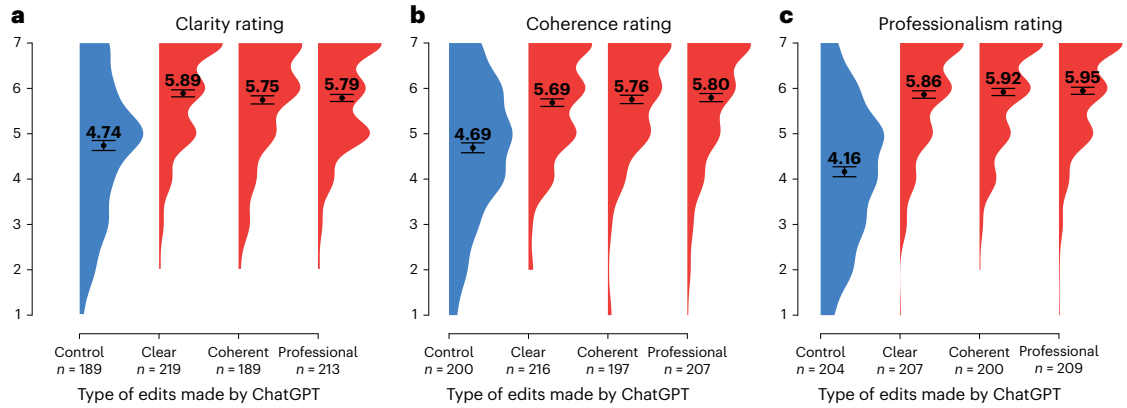

Fig. 3 | Results from Pilot Study 3. a–c, Participants ($N = 490$) were randomly assigned to evaluate one of three linguistic features (clarity, coherence or professionalism) of five different complaints drawn from the pool of 80 complaints (20 Control, 20 Clear, 20 Coherent and 20 Professional complaints). The five complaints shown to each participant always differed in content such that no participant ever evaluated different versions of the same underlying complaint. The mean (dot) and its corresponding standard error (error bars) are shown for each distribution of ratings. Below each violin plot, the *n* value indicates the number of ratings each edit type received, which varied due to random assignment of complaints to participants. For each rating type (clarity, coherence and professionalism), all three groups of LLM-edited complaints (Clear, Coherent and Professional) were rated higher on the focal linguistic feature than the unedited complaints. Three separate linear fixed-effects models controlling for participant fixed effects and complaint (content) fixed effects confirmed these enhancements in presentation (all $t > 8.22$, all Holm-adjusted two-sided $P < 0.001$. Thus, even though each participant evaluated only one linguistic feature, LLM editing improved the overall presentation quality of complaints across multiple linguistic dimensions simultaneously.

The procedure otherwise mirrored that of Experiments 1 and 2, and the participants again had prior work experience in the finance sector (as in Experiment 2). This design allowed us to test the robustness of our findings under stricter control of the experimental stimuli.

As preregistered (https://aspredicted.org/zczq-gmvn.pdf), we analysed the data using linear fixed-effects models. The results were consistent with our hypothesis: complaints edited with ChatGPT were significantly more likely to receive (hypothetical) monetary compensation than unedited complaints ($\beta = 0.17$; 95% CI, (0.035, 0.313); s.e. = 0.071; $t_{1574} = 2.45$; two-sided $P = 0.014$; see Model 2 in Supplementary Table 8). This positive effect of LLM editing on compensation likelihood was consistently observed across different model specifications, as long as the models included fixed effects for complaint content (see Models 3–5 in Supplementary Table 8). These results replicate the positive effect of LLM editing observed in Experiments 1 and 2, providing further evidence that LLM-edited complaints are more likely to secure monetary compensation.

Discussion

Large language models are rapidly transforming communication, yet their real-world efficacy in high-stakes consumer–firm interactions has remained largely unexamined. This paper provides a large-scale empirical analysis of both the adoption and impact of LLMs in consumer advocacy, using over one million financial complaints submitted to the CFPB. Our analysis documents a sharp rise in LLM-assisted complaints following the release of ChatGPT. We further provide causal evidence that using an LLM significantly increases a consumer's likelihood of obtaining monetary or non-monetary relief from firms. A series of controlled experiments reveals one potential mechanism driving this effect: a presentation effect, whereby LLMs enhance linguistic features of a message (for example, clarity, coherence and professionalism) without altering its substantive factual content. This finding aligns with classic persuasion theory and further suggests that LLM-assisted writing can operate through the peripheral route of persuasion by improving surface-level linguistic cues that shape evaluators' judgements, independent of the core arguments.

In addition, our finding speaks to the question of for whom this technology is most beneficial. The second stage of our IV analysis revealed a statistically significant and negative coefficient on the residual term. This provides evidence of negative selection, indicating that LLMs are disproportionately adopted by consumers who possess unobserved characteristics that would otherwise make them less likely to obtain relief. In short, LLMs appear to function as an equalizer, providing the largest persuasive gains to those individuals who may have struggled the most to articulate a compelling case on their own. This mirrors recent labour market evidence where generative AI acts as a skill leveller for novice workers[36,37], extending the scope of this technological levelling effect from workplace productivity to consumer advocacy and regulatory disputes.

The evidence that LLMs act as an equalizer offers empirical support for policies that promote equitable access to this technology. Our findings support the following concrete actions. First, to level the playing field, regulatory bodies such as the CFPB could leverage our insights by integrating structured, easy-to-use LLM assistance directly into their complaint submission portals. This would be particularly beneficial for underserved populations who may lack the resources or technical skills to use these tools independently. Second, our experimental finding that enhancement in presentation alone can impact complaints' outcomes suggests a potential need for regulatory oversight. For example, agencies might audit firm response patterns to ensure that complaints with similar factual content receive consistent treatment, irrespective of their merit in presentation. This would help mitigate the risk that unassisted, less-polished complaints from vulnerable consumers are unfairly dismissed.

Several limitations of the CFPB complaint data warrant emphasis. First, the public records do not identify the specific complaint handler or internal team reviewing each case. This prevents us from directly examining how firm-side beliefs about LLM-assisted complaints evolve over time. Second, the dataset lacks a persistent consumer identifier, limiting our ability to link complaints over time. Consequently, we cannot analyse within-consumer dynamics (for example, whether prior relief outcomes influence the subsequent adoption of LLM assistance). Third, our analyses are restricted to complaints where consumers chose to publicize their narrative; text in complaints withheld from public release is unobservable and therefore excluded from our analysis. These constraints primarily limit the extent to which we can speak to firm-side adaptation, within-consumer dynamics and disclosure decisions; however, they do not overturn the core finding that—conditional

Table 4 | Results from Experiment 1

	Dependent variable: compensation likelihood			
	Model 1	**Model 2**	**Model 3**	**Model 4**
Intercept	3.645			
	(3.441, 3.849)			
	t_{1503}=35.07			
	P < 0.001			
Edited (composite)	0.274	0.306	0.308	
	(0.037, 0.510)	(0.135, 0.476)	(0.137, 0.478)	
	t_{1503}=2.27	t_{1184}=3.52	t_{1180}=3.54	
	P = 0.023	P < 0.001	P < 0.001	
Edited (clarity)				0.261
				(0.051, 0.472)
				t_{1178}=2.44
				P = 0.017
Edited (coherence)				0.279
				(0.071, 0.487)
				t_{1178}=2.63
				P = 0.017
Edited (professionalism)				0.383
				(0.174, 0.593)
				t_{1178}=3.59
				P = 0.001
Complaint content fixed effects	No	Yes	Yes	Yes
Participant fixed effects	No	Yes	Yes	Yes
Order effects	No	No	Yes	Yes
Observations	1,505	1,505	1,505	1,505

This table reports the results from Experiment 1 (N=301 participants; 1,505 observations). The dependent variable is the likelihood of offering monetary compensation measured on a seven-point scale (from 1 (very unlikely) to 7 (very likely)). The 'Edited (composite)' variable is a dummy taking the value 1 if the complaint was improved by ChatGPT (for clarity, coherence or professionalism) and 0 otherwise (unedited). Model 1 estimates the baseline treatment effect without fixed effects, and Model 2 represents the preregistered main regression specification with linear fixed effects. Model 3 assesses robustness by controlling for presentation order, and Model 4 disaggregates the treatment effect by the specific type of linguistic improvement. The values in parentheses represent 95% CIs. All P values are two-sided; the P values in Model 4 are adjusted for multiple comparisons using the Holm–Bonferroni method.

on observed, public complaints—LLM assistance increases the probability of receiving relief.

This research highlights several critical avenues for future investigation. First, the long-term dynamics of LLM adoption warrant deeper exploration. As our data conclude in early 2024, future work may track the evolving adoption trajectory, comparing it with the diffusion patterns of past transformative technologies, such as the personal computer or the Internet. This should include an examination of the extensive margin—whether the availability of LLMs mobilizes consumers who would otherwise not have submitted complaints—and the identification of systematic factors that may deter adoption beyond those studied here. Second, it is crucial to examine the full scope of how consumers interact with AI as a communication partner. Understanding the consumer adoption funnel—whether LLMs are used primarily as search engines for acquiring financial knowledge, tools for revising existing drafts or platforms for fully automating text generation—is essential, though observationally challenging. Furthermore, the organizational response to AI-mediated communication remains an open question. While firms were unlikely to systematically employ detection programs during this early adoption period, they will probably adapt their policies as LLM usage becomes ubiquitous. Future research must explore how firms interpret LLM-assisted messages, particularly in light of AI user aversion—a phenomenon in which the use of AI causes message receivers to devalue the sender's intentions or concerns[38]. This poses a risk that firms might discount LLM-assisted complaints. However, given our evidence of negative selection, dismissing LLM-assisted complaints would disproportionately harm vulnerable populations. Instead, firms should view these complaints as an opportunity to hear grievances from underserved customers who are finally able to voice their concerns effectively. Lastly, future studies may investigate whether improved communication clarity ultimately allows firms to focus more effectively on the substantive merits of disputes, and analyse the role of intermediaries, such as legal representatives (that is, 'lawyering up'), in leveraging these tools for consumer advocacy.

In sum, this research demonstrates that LLMs are not merely tools for linguistic refinement; they are reshaping power dynamics in consumer advocacy. By enhancing the presentation of a grievance, LLMs can significantly improve a consumer's chances of success, with substantial benefits accruing to those who were previously disadvantaged. Fostering equitable access to these tools is therefore not just a matter of technological adoption but a crucial step towards a fairer and more effective system of consumer redress.

Methods

Observational studies

CFPB dataset. The CFPB complaint dataset provides granular information on consumer grievances related to a wide spectrum of financial products and services. A key feature of this dataset is the inclusion of detailed consumer narratives, where individuals describe their

experiences in their own words. Unlike informal and often brief reviews on platforms such as Yelp or Amazon, complaints submitted to the CFPB are processed through a formal government channel, positioning them as official documents with the potential to trigger regulatory action or investigations. Given this regulatory context, consumers are likely to craft their complaints with greater care and precision, as these submissions can carry serious implications for firms. This high-stakes nature is reinforced by guidance from regulatory bodies such as the CFPB[39] and the Federal Trade Commission[40], as well as advocacy groups such as the National Consumer Law Center[41], which advise consumers to treat the process as a formal, evidence-based communication. Consistent with this guidance, our empirical analysis confirms that CFPB complaints are indeed crafted with significantly greater deliberation. A comparative linguistic analysis of 50,000 CFPB complaints against a similarly sized sample of Yelp reviews revealed that CFPB complaints are written at a substantially higher complexity level (average ninth-grade reading level (Flesch–Kincaid Grade, 9.69) versus fifth-grade for Yelp (Grade, 5.02)) and feature significantly longer sentences (21.8 versus 15.5 words, two-sided $P < 0.001$). This linguistic rigour stands in stark contrast to the opinion-oriented, casual nature of reviews on commercial platforms. Moreover, the CFPB's regulatory framework mandates that firms respond to complaints, offering consumers a clear path to potential redress, which may include refunds, corrections or other forms of relief depending on the outcome of the resolution process.

Beyond the consumer narratives, the dataset includes a comprehensive set of structured variables that provide additional insights into the nature and resolution of the complaints. Each complaint is systematically categorized by issue type (for example, "Improper use of your report" or "Incorrect information on your report") and product type (for example, "Debt collection" or "Credit reporting"). The dataset also records firm responses, categorized as "Closed with monetary relief", "Closed with non-monetary relief" or "Closed with explanation", offering a clear view of how firms address consumer grievances. In our analysis, we defined a successful complaint outcome as one that resulted in either monetary or non-monetary relief, consistent with prior literature[42]. This definition is particularly valid in the financial context, where distinctions between monetary and non-monetary relief are nuanced. Although not immediately quantifiable as direct financial compensation, non-monetary relief often yields substantial economic benefits. For example, foreclosure alternatives such as loan modification or forbearance provide considerable economic value by preventing home loss and associated displacement costs. Detailed descriptions of all variables are provided in Supplementary Information section 1A. Moreover, the complaints are geocoded by zip code level, allowing for the identification of regional patterns in consumer dissatisfaction across the USA. The zip code data further enable the integration of sociodemographic information from the ACS, facilitating a more granular analysis of how LLM adoption varies across different demographic and socio-economic groups.

ACS dataset. We used the 2017–2021 ACS five-year estimates to collect representative sociodemographic information at the zip code level to examine the heterogeneous adoption of LLMs following the release of ChatGPT in November 2022. For each zip code, we included key variables such as employment rate, education level, median income, total number of households, language-barrier proxy and Internet access. These variables were chosen to capture cross-sectional variation in sociodemographic characteristics across different regions. While a detailed explanation of each variable's calculation is provided in Supplementary Information section 1B, we followed the Census Bureau's methodology for generating regional profiles.

AI-detection tool. To identify complaints probably written with the assistance of LLMs, we employed the paid premium version of a commercial AI-detection tool that has demonstrated high accuracy to compute an AI Score for each complaint in the CFPB dataset. It is noteworthy that commercialized, paid versions of AI-detection tools exhibit substantially superior performance compared with many free alternatives[43,44]. To assess the robustness of our detection results, we compared the results of this detection program with those of another leading paid AI-detection tool. We observed a high degree of agreement between the two AI-detection tools, with an agreement rate of 94.1%. These results are provided in Supplementary Information section 1C. The detection program calculates a 'Human Score,' which estimates the probability that a given text was written by a human, as opposed to being generated by LLMs such as ChatGPT, GPT-4, Claude or Gemini. For the purposes of our analysis, we defined the AI Score as 100 − Human Score, representing the estimated probability (%) that a complaint was generated with AI assistance. Further details on the tool's methodology and validity, including the results of its accuracy tests, are provided in Supplementary Information section 1C. These results are consistent with recent observations on the high performance of AI-detection programs[45]. In the subsequent analysis, we used the AI Score to examine the prevalence of complaints probably written with LLM assistance and to analyse the relationship between LLM usage and complaint outcomes.

Estimation model. To test the impact of our IVs on LLM usage, we estimated the following first-stage regression:

$$\ln\left(\frac{P(\mathrm{LLM}_i = 1)}{P(\mathrm{LLM}_i = 0)}\right) = \alpha_1 \mathrm{Internet}_{j_i} + \alpha_2 \mathrm{LanguageBarrier}_{j_i} + \mathbf{X}'_{j_i}\boldsymbol{\gamma} + \delta_{t_i} + \delta_{s_i} + \delta_{p_i} + \delta_{k_i}. \quad (1)$$

For complaints submitted after the release of ChatGPT, this logistic regression models the log-odds that a given complaint i was generated using an LLM (where LLM_i is a dummy variable taking the value 1 if the complaint was generated by an LLM and 0 otherwise); the model uses Internet access and the language-barrier proxy as the primary independent variables, measured at the zip code level (j_i) from 2017 to 2021. To address potential confounding, the vector $\mathbf{X}_{j_i}$ includes zip-code-level sociodemographic factors: median income, educational attainment (the proportion of residents with a bachelor's degree or higher), employment rate and total household count. The model further incorporates fixed effects to account for unobserved heterogeneity across multiple complaint dimensions. Specifically, the year–month fixed effects δ_{t_i} control for broad temporal trends in both LLM adoption and relief decisions, encompassing major macroeconomic shocks such as the post-COVID-19 environment and rising inflation. For instance, these fixed effects account for the aggregate increase in consumer complaint filings observed beginning around January 2022. By absorbing these aggregate trends, the year–month fixed effects mitigate the risk of spurious regression due to non-stationarity, ensuring that our identification relies on cross-sectional variation. Meanwhile, the state-level fixed effects δ_{s_i} capture cross-state heterogeneity, the product-category fixed effects δ_{p_i} account for different financial products (for example, credit cards and mortgages) and the issue-type fixed effects δ_{k_i} control for variation in complaint issue types (for example, billing disputes and fraud). These fixed effects help isolate the exogenous influence of the IVs on LLM usage by controlling for heterogeneity across multiple dimensions.

Leveraging exogenous variation from our two IVs in the first stage, we adopted the 2SRI method outlined by Terza et al.[46] and Wooldridge[47]. This approach, designed for models where both the endogenous regressor and the final outcome can be binary and estimated via a logistic link, uses the residual from the first stage as a control function—hereafter, the first-stage residual (for brevity, 'residual'). Specifically, we calculated standardized residuals ($\hat{r}_i \equiv \frac{\mathrm{LLM}_i - \hat{P}(\mathrm{LLM}_i=1)}{\sqrt{\hat{P}(\mathrm{LLM}_i=1)(1-\hat{P}(\mathrm{LLM}_i=1))}}$),

following the idea that this can spread out the residual value to capture the rare treatment[48]. We then included $\hat{r}_i$ in the second-stage logistic regression, as specified in equation (2), to correct for the endogeneity of LLM usage. To further control for the semantic content of each complaint, we incorporated $\mathbf{w}'$, a 384-dimensional vector of sentence embeddings, into the regression. Specifically, to represent the content of each complaint, we employed sentence embeddings generated by the `all-MiniLM-L6-v2` model from the Sentence-Transformers framework[49]. We selected this model for its compact 384-dimensional output, which provides an effective yet computationally efficient method for capturing the semantic meaning of the complaints. This allowed us to account for the narrative's meaning beyond the fixed-effect categories from products or issues. To account for the estimation uncertainty from the first-stage regression, we employed a bootstrap procedure with 1,000 replications to estimate standard errors:

$$\ln\left(\frac{P(\text{Relief}_i = 1)}{P(\text{Relief}_i = 0)}\right) = \beta_1 \text{LLM}_i + \beta_2 \hat{r}_i + \mathbf{X}'_{j_i}\gamma + \delta_{t_i} + \delta_{s_i} + \delta_{p_i} + \delta_{k_i} + \mathbf{w}'\theta \quad (2)$$

We used standardized residuals to capture the rare treatment effectively (that is, relatively few Likely-AI complaints)[48]. Given the challenge of determining a priori which residual specification from a binary treatment model best captures the non-separable error term in the outcome equation, researchers often consider alternative specifications such as the raw residual, $\hat{v}_i \equiv \text{LLM}_i - \hat{P}(\text{LLM}_i = 1)$ (equivalent to the generalized residual under a logit model), or the deviance residual[46,47,50]. We assessed the robustness of our findings to these specifications and to the use of a linear probability model in lieu of logit, consistently confirming that LLM usage increases the likelihood of relief. Further details are provided in Supplementary Information section 1F.

To further substantiate our findings, we conducted a few additional empirical analyses. First, we addressed the econometric requirement that exogenous controls included in the second stage (here, the sentence embeddings) should generally also be included in the first stage for consistent 2SRI estimation. While our main specification excludes these high-dimensional embedding vectors from the first stage to mitigate the practical risks of overfitting and multicollinearity, we implemented an alternative, symmetric specification using principal component analysis to manage the dimensionality of embedding vectors and included these principal component analysis outputs in both stages (that is, the same set of control variables). The results were highly consistent with our main findings, confirming a significant positive treatment effect (7.9 percentage points in the second-stage linear probability model) and evidence of negative selection. Second, we compared complaint outcomes submitted both before and after the release of ChatGPT, focusing on regions that eventually adopted LLMs (treatment group) versus those that never did (control group) in a difference-in-differences framework. The interaction term between the post-ChatGPT period and eventual-adopter region is positive and statistically significant ($\beta = 0.057$; 95% CI, (0.051, 0.063); z=19.00; two-sided $P < 0.001$). These results support the main conclusion of our study and are detailed in Supplementary Information section 1F.

Experimental studies

The experimental studies were approved (that is, deemed exempt) by the Institutional Review Board at Northeastern University (IRB no. 23-12-14) and were conducted in accordance with the relevant guidelines and regulations. All participants provided informed consent prior to participation.

Preregistrations. For Experiments 1 and 2, we report the hypotheses, outcome variables and analyses as preregistered. For Pilot Study 1, we highlight the results from one preregistered model (Model 4), but all other preregistered models showed the same pattern of results and statistical significance (all $P < 0.001$). These results are reported in detail in Supplementary Information section 7A. For Pilot Study 2, we did not create a preregistration. For Pilot Study 3, we discuss the results from all preregistered analyses, but these analyses are reported in detail in Supplementary Information section 7C. For Experiment 3, we report the first of the two preregistered analyses and the first of the two preregistered hypotheses in the main text. The second preregistered analysis and hypothesis are presented and discussed in detail in Supplementary Information section 7F. For Experiment 3, we did not preregister any hypothesis about the length of complaints. However, when we created the new set of stimuli for Experiment 3, we ensured that LLM-edited complaints were not significantly longer than unedited complaints (as was the case in Experiments 1 and 2); see Supplementary Information section 3C.

Sample and procedure. For the controlled experiments, participants were recruited from Prolific and randomly assigned to conditions. No statistical methods were used to predetermine sample sizes. However, our sample sizes (N = 200–500) yielded thousands of rating-level observations and precise estimates of the focal effects, and they are within the range of sample sizes typically used in comparable experimental studies[51,52]. While the participants were blinded to the specific nature of the stimuli (that is, they were not informed which complaints were LLM-edited), data collection and analysis were not performed blind to the conditions of the experiments.

All experiments and two of the three pilot studies followed the same general procedure. Specifically, the participants (1) were asked to imagine they were in charge of handling consumer complaints, (2) read a series of complaints (all featuring different core contents) randomly drawn from the given study's pool of complaints and then (3) rated the likelihood of offering monetary compensation in response to each complaint. We elaborate on the procedures for each pilot study and experiment in detail in Supplementary Information sections 2 and 3.

Inclusion and ethics

This research analyses publicly available, de-identified data from the CFPB Consumer Complaint Database and publicly available aggregate demographic data from the ACS. The experimental portion of this research (online controlled experiments) was reviewed and deemed exempt by the Institutional Review Board at Northeastern University (IRB no. 23-12-14) and was conducted in accordance with all relevant ethical guidelines. All participants provided informed consent prior to participation.

Software

To analyse the CFPB complaint data, we used the following software and R packages: R (version 4.3.2)[53], data.table (version 1.14.10)[54], ggplot2 (version 3.4.4)[55], gridExtra (version 2.3)[56], dplyr (version 1.1.3)[57], scales (version 1.3.0)[58], bife (version 0.7.2)[59] and plm (version 2.4.3)[60]. To analyse the data from the experiments, we used the following software and R packages: R (version 4.5.0), data.table (version 1.17.4)[54], ggplot2 (version 4.0.1)[55], kim (version 0.6.3)[61], lfe (version 3.1.1)[62] and see (version 0.12.0)[63].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The original data on consumer complaints can be downloaded from the CFPB's Consumer Complaint Database (https://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data). The datasets for the observational and experimental studies in this research are available via the project's Open Science Framework page at https://osf.io/fh2cz/. Source data are provided with this paper.

Code availability

The code for reproducing the analyses for both the observational and experimental studies in this research can be accessed via the project's Open Science Framework page at https://osf.io/fh2cz/.

References

1. Liang, W. et al. Monitoring AI-modified content at scale: a case study on the impact of ChatGPT on AI conference peer reviews. In *International Conference on Machine Learning* (eds Salakhutdinov, R. et al.) 29575–29620 (PMLR, 2024).
2. Liang, W., Zhang, Y., Wu, Z. et al. Quantifying large language model usage in scientific papers. *Nat. Hum. Behav.* **9**, 2599–2609 (2025); https://doi.org/10.1038/s41562-025-02273-8
3. Liang, W. et al. The widespread adoption of large language model-assisted writing across society. *Patterns* **6**, 101366 (2025).
4. Ling, Y., Kale, A. & Imas, A. Underreporting of AI use: the role of social desirability bias. SSRN Working Paper https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5464215 (2025).
5. Humlum, A. & Vestergaard, E. The unequal adoption of ChatGPT exacerbates existing inequalities among workers. *Proc. Natl Acad. Sci. USA* **122**, e2414972121 (2025).
6. Park, E. & Gelles-Watnick, R. *Most Americans Haven't Used ChatGPT; Few Think It Will Have a Major Impact on Their Job* (Pew Research Center, 2023); https://pewrsr.ch/3SCeWX9
7. McClain, C. *Americans' Use of ChatGPT Is Ticking Up, but Few Trust Its Election Information* (Pew Research Center, 2024); https://shorturl.at/nn49B
8. Bick, A., Blandin, A. & Deming, D. J. *The Rapid Adoption of Generative AI* Working Paper No. 32966 (National Bureau of Economic Research, 2024); http://www.nber.org/papers/w32966
9. Wu, Y. et al. Autoformalization with large language models. *Adv. Neural Inf. Process. Syst.* **35**, 32353–32368 (2022).
10. Saito, K., Wachi, A., Wataoka, K. & Akimoto, Y. Verbosity bias in preference labeling by large language models. Preprint at https://arxiv.org/abs/2310.10076 (2023).
11. Shahriar, S. & Hayawi, K. Let's have a chat! A conversation with ChatGPT: technology, applications, and limitations. *Artif. Intell. Appl.* **2**, 11–20 (2024).
12. Lusardi, A. & Mitchell, O. S. The economic importance of financial literacy: theory and evidence. *J. Econ. Lit.* **52**, 5–44 (2014).
13. Fernandes, D., Lynch Jr, J. G. & Netemeyer, R. G. Financial literacy, financial education, and downstream financial behaviors. *Manage. Sci.* **60**, 1861–1883 (2014).
14. Russo, G., Horta Ribeiro, M., Davidson, T. R., Veselovsky, V. & West, R. The AI review lottery: widespread AI-assisted peer reviews boost paper scores and acceptance rates. *Proc. ACM Hum.-Comput. Interact.* (ed. Nichols, J.) **9**, CSCW486 (2025).
15. Kadoma, K. et al. The role of inclusion, control, and ownership in workplace AI-mediated communication. In *Proc. 2024 CHI Conf. Hum. Factors Comput. Syst.* (eds Mueller, F. F. et al.) 1016 (2024).
16. Wiles, E. & Horton, J. J. Generative AI and labor market matching efficiency. SSRN Working Paper https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5187344 (2025).
17. Wester, J., De Jong, S., Pohl, H. & Van Berkel, N. Exploring people's perceptions of LLM-generated advice. *Comput. Hum. Behav. Artif. Hum.* **2**, 100072 (2024).
18. Kadoma, K., Metaxa, D. & Naaman, M. Generative AI and perceptual harms: who's suspected of using LLMs? In *Proc. 2025 CHI Conference on Human Factors in Computing Systems* (eds Yamashita, N. et al.) 861 (2025).
19. Hermann, E., Williams, G. Y. & Puntoni, S. Deploying artificial intelligence in services to aid vulnerable consumers. *J. Acad. Market. Sci.* **52**, 1431–1451 (2024).
20. Reif, E. et al. A recipe for arbitrary text style transfer with large language models. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics* (eds Muresan, S. et al.) **2**, 837–848 (Association for Computational Linguistics, 2022).
21. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl Acad. Sci. USA* **117**, 30046–30054 (2020).
22. Liu, A. X., Xie, Y. & Zhang, J. It's not just what you say, but how you say it: the effect of language style matching on perceived quality of consumer reviews. *J. Interact. Market.* **46**, 70–86 (2019).
23. Chen, J., Fan, W., Wei, J. & Liu, Z. Effects of linguistic style on persuasiveness of word-of-mouth messages with anonymous vs. identifiable sources. *Market. Lett.* **33**, 593–605 (2022).
24. Petty, R. E. & Cacioppo, J. T. The elaboration likelihood model of persuasion. In *Advances in Experimental Social Psychology* (ed. Berkowitz, L.) 123–205 (Academic Press, 1986).
25. Alter, A. L. & Oppenheimer, D. M. Uniting the tribes of fluency to form a metacognitive nation. *Pers. Soc. Psychol. Rev.* **13**, 219–235 (2009).
26. Rogers, E. M., Singhal, A. & Quinlan, M. M. Diffusion of innovations. In *An Integrated Approach to Communication Theory and Research* (eds Stacks, D. W. et al.) 432–448 (Routledge, 2014).
27. Card, D. in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp* (eds Christofides, L. N. et al.) 201–222 (Univ. Toronto Press, 1995).
28. Chetty, R., Hendren, N. & Katz, L. F. The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment. *Am. Econ. Rev.* **106**, 855–902 (2016).
29. Grewal, R. & Orhun, Y. Unpacking the instrumental variables approach. *Impact at JMR* https://www.ama.org/marketing-news/unpacking-the-instrumental-variables-approach/ (2024).
30. *Company Portal Manual* version 2.14 (Consumer Financial Protection Bureau, 2015).
31. Ananat, E. O. The wrong side(s) of the tracks: the causal effects of racial segregation on urban poverty and inequality. *Am. Econ. J. Appl. Econ.* **3**, 34–66 (2011).
32. Nunn, N. & Wantchekon, L. The slave trade and the origins of mistrust in Africa. *Am. Econ. Rev.* **101**, 3221–3252 (2011).
33. Lowes, S. & Montero, E. The legacy of colonial medicine in central Africa. *Am. Econ. Rev.* **111**, 1284–1314 (2021).
34. Grice, H. P. Logic and conversation. In *Syntax and Semantics 3: Speech Acts* (ed. Cole, P.) (Academic Press, 1975).
35. Halliday, M. A. K. & Hasan, R. *Cohesion in English* (Routledge, 2014).
36. Noy, S. & Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).
37. Brynjolfsson, E., Li, D. & Raymond, L. Generative AI at work. *Q. J. Econ.* **140**, 889–942 (2025).
38. Dang, J. & Liu, L. Extended artificial intelligence aversion: people deny humanness to artificial intelligence users. *J. Pers. Soc. Psychol.* https://doi.org/10.1037/pspi0000480 (2024).
39. *What's Most Important for Me to Include in a Complaint?* (Consumer Financial Protection Bureau, 2025); https://www.consumerfinance.gov/complaint/
40. Tressler, C. *How to Write an Effective Complaint Letter* (Federal Trade Commission Consumer Advice, 2015); https://consumer.ftc.gov/consumer-alerts/2015/09/how-write-effective-complaint-letter
41. *How to File a Complaint with the Consumer Financial Protection Bureau (CFPB) about Credit Repair* (National Consumer Law Center, 2023); https://www.nclc.org/resources/how-to-file-a-complaint-with-the-consumer-financial-protection-bureau-cfpb-about-credit-repair/

42. Dou, Y., Hung, M., She, G. & Wang, L. L. Learning from peers: evidence from disclosure of consumer complaints. *J. Account. Econ.* **77**, 101620 (2024).
43. Luo, L. & Ma, L. Wisdom of the AI crowd? Can we detect AI-generated product reviews? *SSRN* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4511025 (2024).
44. Jabarian, B. & Imas, A. *Artificial Writing and Automated Detection* Working Paper (National Bureau of Economic Research, 2025).
45. Prillaman, M. 'ChatGPT detector' catches AI-generated papers with unprecedented accuracy. *Nature* https://doi.org/10.1038/d41586-023-03479-4 (2023).
46. Terza, J. V., Basu, A. & Rathouz, P. J. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J. Health Econ.* **27**, 531–543 (2008).
47. Wooldridge, J. M. Control function methods in applied econometrics. *J. Hum. Resour.* **50**, 420–445 (2015).
48. Basu, A., Coe, N. B. & Chapman, C. G. 2SLS versus 2SRI: appropriate methods for rare outcomes and/or rare exposures. *Health Econ.* **27**, 937–955 (2018).
49. Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (eds Inui, K. et al.) 3982–3992 (Association for Computational Linguistics, 2019); https://aclanthology.org/D19-1410
50. Song, H., Tucker, A. L., Graue, R., Moravick, S. & Yang, J. J. Capacity pooling in hospitals: the hidden consequences of off-service placement. *Manage. Sci.* **66**, 3825–3842 (2020).
51. Matz, S.C., Teeny, J.D., Vaid, S.S. et al. The potential of generative AI for personalized persuasion at scale. *Sci. Rep.* **14**, 4692 (2024); https://doi.org/10.1038/s41598-024-53755-0
52. Salvi, F., Horta Ribeiro, M. & Gallotti, R. et al. On the conversational persuasiveness of GPT-4. *Nat. Hum. Behav.* **9**, 1645–1653 (2025).
53. R Core Team. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, 2023); https://www.R-project.org/
54. Barrett, T. et al. data.table: Extension of data.frame. R package https://cran.r-project.org/package=data.table (2023).
55. Wickham, H. et al. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package https://cran.r-project.org/web/packages/ggplot2/index.html (2023).
56. Auguie, B. gridExtra: Miscellaneous functions for 'grid' graphics. R package version 2.3 https://cran.r-project.org/package=gridExtra (2017).
57. Wickham, H. et al. dplyr: A Grammar of Data Manipulation. R package version 1.1.3 https://cran.r-project.org/web/packages/dplyr/index.html (2023).
58. Wickham, H. & Seidel, D. scales: Scale Functions for Visualization. R package version 1.3.0 https://cran.r-project.org/web/packages/scales/index.html (2023).
59. Stammann, A., Czarnowske, D. & Heiss, F. bife: Binary Choice Models with Fixed Effects. R package version 0.7.2 https://github.com/amrei-stammann/bife (2022).
60. Croissant, Y. & Millo, G. plm: Linear Models for Panel Data. R package version 2.4-3 https://cran.r-project.org/package=plm (2023).
61. Kim, J. kim: A Toolkit for Behavioral Scientists. R package version 0.6.3 https://cran.r-project.org/web/packages/kim/index.html (2025).
62. Gaure, S. & Sepulveda, M. V. lfe: Linear Group Fixed Effects. R package version 3.1.1 https://cran.r-project.org/package=lfe (2025).
63. Lüdecke, D. et al. see: Model Visualisation Toolbox for 'easystats'and 'ggplot2'. R package version 0.12.0 https://cran.r-project.org/web/packages/see/index.html (2025).

Acknowledgements

We thank the Department of Marketing at the City University of Hong Kong for financial support (grant ID 7200769 to M.S.). The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We also thank Y. Li, L. Guo, K. Uetake, S. Ghili, C. Wang, H. Fong, T. Chan, Y. Chen, O. Urminsky, T. L. Griffiths, G. Zauberman, A. Dukes, K. Sudhir, P. Arcidiacono, C. Fuchs, G. Packard, S. Puntoni, J. Joo, S. Ma, H. Park, L. Su and C. Song for their constructive feedback, as well as seminar participants at the Annual Business and Generative AI Workshop by Wharton, the Symposium on AI in Marketing, the Fisher AI in Business Conference, the China Marketing Science Conference, the AIM Conference, the AI ML and Business Analytics Conference, the Hong Kong Joint School Marketing Conference, Korea University, KAIST, CKGSB, the Informs Marketing Science Conference and the Marketing Exchange Forum.

Author contributions

M.S. designed the research question. M.S., J.K. and J.S. developed the empirical strategy. M.S. collected the data and estimated all regression models in the observational studies. M.S. and J.K. designed the experimental studies. J.K. conducted all statistical analyses in the experimental studies. M.S., J.K. and J.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-026-02409-4.

Correspondence and requests for materials should be addressed to Minkyu Shin, Jin Kim or Jiwoong Shin.

Peer review information *Nature Human Behaviour* thanks Rajdeep Grewal and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.