Silence or Theatrics: Local Cadre Presence and Political Response in Survey

Working Paper*

LI, Jiajun †

March 5, 2025

Contents

1 Introduction 3

2 Review 6

3 Research Design 6

4 Results 9

4.1 Cadre Presence and Social Desirability 9

4.2 Cadre Presence and Self Censorship 12

4.3 Placebo Test 12

4.4 Simulation Experiment with Deepseek-R1 15

5 Conclusion 25

A Appendix 25

---

*This is a draft working paper, and certain sections remain under development. I particularly encourage readers focus on the Design and Results sections, which are central to this research. For optimal clarity and formatting, I highly recommend reviewing the latest PDF version available on my website https://github.com/JiajunLeoLi/local-cadre-project.

†Email:jiajun.li@connect.ust.hk, Division of Social Science, Hong Kong University of Science and Technology.

Abstract

Most social surveys in China necessitate local government approval and cooperation from local cadres. This study investigates whether this approach introduces bias, specifically examining whether the presence of local cadres during survey interviews leads to self-censorship or social desirability bias. Analyzing data from the China Family Panel Studies (2012–2018) using two-way fixed effects models, we find that respondents in the presence of local cadres report higher levels of trust in both local cadres and government. Concurrently, they are less likely to mention experiences of unfairness, unreasonable treatment, or conflict with government agencies. Surprisingly, partial evidence suggests that local cadre presence encourages respondents to answer questions related to both political and non-political issues, which contradicts previous assertions regarding self-censorship in authoritarian regimes. In addition to conducting balance checks, subsample analyses, and placebo tests (e.g., neighbor presence), we implement a simulation experiment utilizing the advanced reasoning large language model (LLM) Deepseek-R1 (671B version), which yields consistent estimations. Furthermore, by employing the word embedding method based on the thinking content of AI agents, we construct an authoritarian perception index and perform two-stage least squares regressions. Our results indicate that authoritarian perception is a significant mechanism through which social desirability bias operates. In conclusion, our findings reveal that local cadre presence during survey interviews can induce social desirability bias rather than self-censorship in China. This underscores the need for scholars to carefully balance government cooperation with data integrity in their research.

Keywords: social desirability, self-censorship, political trust, simulation experiment, large language model

1 Introduction

Most social surveys conducted in China necessitate approval from local government entities and the cooperation of local cadres. However, it remains uncertain whether this pragmatic approach introduces biases such as self-censorship and social desirability in the context of the world's most populous authoritarian regime. As the most immediate representatives of the government, local cadres accompanying survey interviews inevitably exert social pressure and create a perception of national surveillance among respondents(Fitzgerald, 2021; Iida, 2020). This dynamic may lead to inauthentic behavior and dishonest responses, raising significant concerns about the validity of social survey data(Belli et al., 1999; Burnett et al., 2022).

Methodologically, although numerous studies have examined these issues using online data and vignette experiments (Blair et al., 2020; Nicholson and Huang, 2023; Singh and Tir, 2023), such methods often overlook the noisy but potentially confounding factors inherent in actual survey practices. This study intend to address the following research questions: (1) Does the presence of local cadres during surveys lead to self-censorship and social desirability bias among respondents? (2) Do these effects vary across different types of survey questions? (3) How does self-censorship or social desirability operate this cognitive process?

Analyzing data from the China Family Panel Studies (2012–2018) with two-way fixed effects models, we find that local cadre presence during interviews leads respondents to report higher trust in local cadres and government, while reducing probabilities of reports of perceived unfairness or administrative conflicts. Surprisingly, partial evidence shows that local cadre presence encourages respondents to answer all types of questions (political trust, non-political trust, unfairness, and conflict with government agencies), contradicting previous claims of self-censorship in authoritarian regimes. In sum, our results support the relationship between local cadre presence and social desirability rather than self-censorship.

To address potential selection bias related to the presence of local cadres, a series of follow-up analyses were conducted to bolster the study's causal claims. First, balance checks revealed no systematic differences between the treated and control groups. Second, placebo tests involving the presence of neighbors indicated that the observed effects are specifically attributable to the presence of local cadres, rather than the presence of other individuals. Additionally, restricting the sample to individuals who experienced the presence of local cadres in one or more survey waves effectively mitigated potential selection bias, with all estimations remaining robust.

Furthermore, in an innovative approach, we employed the currently most popular large language model Deepseek-R1 (671B version) to conduct a simulation experiment. In this experiment, AI agents, prompted by randomly assigned demographic factors, responded to similar questions from the CFPS under randomly flip contexts (with and without local cadre presence). The direction and statistical significance of the coefficients derived from the simulated data were consistent with those from the previous analyses.

Moreover, leveraging the thinking content of Deepseek-R1, we constructed an authoritarian perception index by calculating the co-occurrence of the term "cadre" with a list of keywords associated with social desirability, such as "pressure" and "fear." The estimations demonstrated a strong correlation between the authoritarian perception index and the responses of the AI agents, highlighting the mechanisms through which social desirability bias operates.

This study contributes to several important strands of literature. Firstly, it deepens our understanding of self-censorship in authoritarian regimes. Unlike the subconscious self-censorship driven by fear of moral and state punishment on the social media platforms(Bar-Tal, 2017; Chang and Manion, 2021), our empirical evidence highlights how the presence of local cadres suppresses respondents' intentions to refuse to answer sensitive political questions in surveys. Inversely, it encourages interviewees to provide answers, even if those responses may be over- or under-reported.

Secondly, this study provides new insights into the sources of social desirability bias. While various studies have explored the causes of this well-documented psychological phenomenon—such as partisanship, group pressure, and moral imperatives(Connors, 2023; DeBell et al., 2020; Engelhardt, 2023; Urbatsch, 2020)—our findings emphasize the unique role of local cadres as a significant source of bias in survey responses. Specifically, perceptions of authoritarianism skew respondents' answers toward politically acceptable norms in sensitive contexts, as evidenced by the disparity between responses to positively and negatively framed questions.

Finally, this study appeal to the critical challenge of accommodating government cooperation while ensuring data reliability. It reveals the tension between the necessity of local government involvement in data collection and the potential compromise of data integrity. By identifying the biases introduced by the presence of local cadres, this research challenges commonly employed survey strategies and underscores the importance of minimizing government-related bias. It calls for the development of survey implementation schemes that prioritize data integrity while navigating the complexities of government cooperation.

The remaining sections are structured as follows. Section II introduces key concepts (social desirability and self-censorship) and proposes our main hypothesis. Section III describes the research design, including the data source, variables, and estimate approach. In the Section IV, one main part is the results derived from CFPS, another is the simulation experiment with large language model and the mechanism analysis. Section IV summarizes conclusions and limitations.

The remaining sections of this paper are organized as follows. Section II introduces extant insights of our key concepts, specifically social desirability and self-censorship, and presents our main hypothesis. Section III outlines the research design, detailing the data sources, variables, and estimation approach employed in the study. In Section IV, we present two main components: the results derived from the China Family Panel Studies and the findings from a simulation experiment using a large language model, along with an analysis of the underlying mechanisms. Finally, Section V summarizes the conclusions and discusses the limitations.

2 Review

3 Research Design

Data:

The data for this study is drawn from the China Family Panel Studies (CFPS) conducted in the years 2012, 2014, 2016, and 2018.

Dependent Variable:

To measure social desirability, we utilize the following two types of items:

- Positively framed questions: "How much do you trust local cadres/ local government/ your neighborhood/ your parents/ Americans/ strangers/ doctors?" (rated from 1 to 10)
- Negatively framed questions: "Do you experience unfairness due to inequality in personal wealth/ household registration status/ gender discrimination/ conflicts with government agencies/ unreasonable delays or stalling at government agencies/ unreasonable charges paid to government agencies?" (yes=1, no=0)

For exploring self-censorship, we employ similar items, where a response is coded as 1 and missing, refused, or unknown responses are coded as 0.

Independent Variable:

The independent variable is measured by the question: "Besides family members, is a **local cadre** present during the survey interview?" (yes=1, no=0). In the part of placebo tests, we alter the independent variable, which is measured by the question: "Besides family members, is a **neighbor** present during the survey interview?"

Control Variables:

To our best effort, this study controls for the following variables: age, years of education, income, gender, residence, marital status, health condition, and partisan affiliation.

Based on balance checks in Table 1, except column *Urban*, their is no significant difference across control variables. The full sample model suggests that local cadres are more likely to be present during survey interviews with urban respondents. However, when the sample is restricted to individuals who have been exposed to cadre presence in one or more survey waves, no significant differences are observed between the treated and control groups. To strengthen these findings, it is necessary to replicate the same analysis within subsamples. Additionally, conducting separate regressions for urban and rural groups can further mitigate concerns about selection bias.

Estimator:

To further examine the effect of local cadre presence, we estimate the following Equation 1 with different survey responses as outcome variables:

$$\mathbf{Y}_{it} = \alpha + \beta \mathbf{Cadre}_{it} + \vartheta_1' \mathbf{X}_{it} + \lambda_i + \sigma_t + \epsilon_{it} \tag{1}$$

- Y_{it}: outcomes of interest of individual i in year t.
- $Cadre_{it}$: whether a cadre show up in the survey interview of individual i in year t.
- $\mathbf{X}_{it}$: an array of control variables.
- λ_i: individual fixed effect.

Table 1: Balance Check

	(1) Full Sample	(2) Subsample
Age	-0.000	-0.051
	(0.001)	(0.045)
Education Years	0.000	0.007
	(0.000)	(0.013)
Ln(income)	-0.001	-0.012
	(0.000)	(0.012)
Male=1	-0.006	-0.395
	(0.010)	(0.428)
Urban=1	0.003*	0.156
	(0.002)	(0.085)
Marital Status	-0.002	-0.037
	(0.001)	(0.033)
Health Condition	0.000	0.002
	(0.000)	(0.012)
Partisan=1	-0.005	-0.139
	(0.003)	(0.096)
Observations	106152	2799
R^2	0.354	0.130

Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

- σ_t: year fixed effect.

- ϵ_{it}: error term.

4 Results

4.1 Cadre Presence and Social Desirability

Table 2 demonstrates that when a local cadre is present during a survey interview, respondents tend to overreport their political trust in local cadres and local government. However, columns 4 to 7 indicate that the presence of local cadres does not significantly impact respondents' non-political trust in their parents, Americans, strangers, or doctors. Notably, in column 3, respondents report a higher level of trust in their neighbors when local cadres are present, which can be attributed to the fact that local cadres often reside in the same community or neighborhood.

Conversely, Table 3 presents results across seven columns, where the coefficients are negatively significant and range from -0.20 to -0.54. Specifically, when a local cadre appears during a survey interview, interviewees are less likely to report experiences of unfairness related to personal wealth inequality, household registration status, gender discrimination, conflicts with government officials, unreasonable delays or stalling by government agencies, and unjust charges imposed by these agencies.

The corresponding subsample analyses are illustrated in Table A1 and Table A2, which yield consistent results: the presence of local cadres in the survey induces higher levels of political trust rather than non-political ones, and it hinders respondents from reporting unfairness, unreasonable treatment, and conflicts they have encountered in the past.

Table 2: Cadre Presence and (Non-)Political Trust

	Political Trust in ...		Non-political Trust in ...				
	(1) Local Cadre	(2) Local Government	(3) Neighbor	(4) Parent	(5) American	(6) Stranger	(7) Doctor
Cadre Presence	0.741***	0.203***	0.164*	-0.025	-0.071	0.003	0.102
	(0.088)	(0.034)	(0.075)	(0.054)	(0.088)	(0.075)	(0.081)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	101009	97733	101343	101008	98684	101051	101297
R^2	0.557	0.494	0.541	0.488	0.544	0.528	0.534

Cadre is an indicator that equals one for respondents in the context of local cadre presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Cadre Presence and Unfair/Conflict Report

	(1) Unfairness due to inequality of personal wealth	(2) Unfairness due to household registration status	(3) Unfairness due to gender discrimination	(4) Unfairness due to government officials	(5) Conflict with government officials	(6) Unreasonable delay/stalling at government agency	(7) Unreasonable charges paid to government agency
Cadre Presence	-0.041**	-0.021*	-0.023**	-0.043**	-0.020*	-0.054***	-0.035**
	(0.015)	(0.010)	(0.008)	(0.013)	(0.009)	(0.015)	(0.011)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	70582	70505	70899	71744	72089	71781	71837
R^2	0.508	0.491	0.463	0.527	0.491	0.532	0.502

Cadre is an indicator that equals one for respondents in the context of local cadre presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.2 Cadre Presence and Self Censorship

Table 4 indicates that in the context of local cadre presence, respondents are more likely to answer questions regarding trust in local cadres, neighbors, parents, strangers, and doctors. However, this phenomenon does not extend to trust in local government or Americans. The corresponding subsample analysis in Table A3 reveals similar results.

In Table 5, we observe that when local cadres are present, respondents are more likely to address questions regarding unfairness stemming from various causes and unreasonable treatment by government agencies. However, in Table A3, all coefficients shrink and become insignificant, with the exception of columns 2 and 3.

If self-censorship bias was expected, all columns should be significant negative, namely respondents are reluctant to answer and the data points are always missing. However, estimations are opposed to what we want, being positive. In this situation, no empirical evidence supports self-censorship hypothesis. And then, we can find out where the positive effects come from in placebo tests.

If self-censorship bias were expected, we would anticipate all columns to exhibit significantly negative coefficients, indicating that respondents are reluctant to answer and that data points are frequently missing. However, the estimations contradict this expectation, showing positive values instead. Hence, there is no empirical evidence supporting the self-censorship hypothesis. Then, we will intend to claim the source of these positive effects through placebo tests.

4.3 Placebo Test

A prominent concern is that the observed effects may arise from the presence of any individual during the survey interview, regardless of their identity. To address this, we employ “neighbor presence” in our placebo tests.

Table 4: Cadre Presence and Response of (Non-)Political Trust

	Political Trust in ...		Non-political Trust in ...				
	(1) Local Cadre	(2) Local Government	(3) Neighbor	(4) Parent	(5) American	(6) Stranger	(7) Doctor
Cadre Presence	0.022^{**}	0.017	0.022^{***}	0.019^{**}	0.012	0.023^{***}	0.020^{**}
	(0.007)	(0.009)	(0.006)	(0.007)	(0.009)	(0.007)	(0.007)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	106152	106152	106152	106152	106152	106152	106152
R^2	0.447	0.431	0.454	0.447	0.426	0.446	0.452

Cadre is an indicator that equals one for respondents in the context of local cadre presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Cadre Presence and Response of Unfair/Conflict Report

	(1) Unfairness due to inequality of personal wealth	(2) Unfairness due to household registration status	(3) Unfairness due to gender discrimination	(4) Unfairness due to government officials	(5) Conflict with government officials	(6) Unreasonable delay/stalling at government agency	(7) Unreasonable charges paid to government agency
Cadre Presence	0.013	0.021^{*}	0.021^{*}	0.019^{*}	0.013	0.019^{*}	0.019^{*}
	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	76670	76670	76670	76670	76670	76670	76670
R^2	0.492	0.489	0.494	0.497	0.504	0.500	0.502

Cadre is an indicator that equals one for respondents in the context of local cadre presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6 shows that, in the context of neighbor presence, respondents do not exhibit a higher level of trust in any individuals except for their neighbors. Similarly, all columns in Table 7 are statistically insignificant, indicating that there is no increased probability of reporting unfairness, conflict, or unreasonable treatment in the context of neighbor presence. In summary, the differing patterns of local cadre and neighbor presence suggest that the observed effects in Table 2 and Table 3 are attributable to the identity of local cadres rather than alternative explanations.

In contrast, both Table 8 and Table 9 reveal that respondents are more likely to answer all types of questions when considering neighbor presence. These estimations, to some extent, help explain the inconsistencies observed in columns from Table 4 and Table 5 across different question items and their partly positive nature.

From Table A5 to Table A8, we also conduct placebo tests using subsample data. To conserve space, we do not discuss these results here, as they all exhibit similar estimations.

4.4 Simulation Experiment with Deepseek-R1

Despite our rigorous efforts to address selection bias, concerns about unobservable confounders persist. While we have identified the effects of local cadre presence on survey responses, we must exercise caution in attributing these effects solely to social desirability bias. The harsh reality is that researchers cannot be present at every survey interview to monitor respondents' cognitive processes and ensure the authenticity of their responses.

Fortunately, advancements in artificial intelligence have paved the way for innovative methodologies to mitigate these concerns(Manning et al., 2024). There is a growing consensus that large language models (LLMs) can effectively mimic human behaviors and attitudes (Huang et al., 2024; Xie et al., 2025; Yao et al., 2024). In this study, we leverage a simulation experiment utilizing the advanced reasoning model Deepseek-R1 (671B version) (Guo et al.,

Table 6: Placebo Test–Neighbor Presence and (Non-)Political Trust

	Political Trust in ...		Non-political Trust in ...				
	(1) Local Cadre	(2) Local Government	(3) Neighbor	(4) Parent	(5) American	(6) Stranger	(7) Doctor
Neighbor Presence	0.068	0.001	0.276^{***}	-0.014	-0.075	-0.003	0.057
	(0.047)	(0.018)	(0.039)	(0.029)	(0.046)	(0.040)	(0.043)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	101009	97733	101343	101008	98684	101051	101297
R^2	0.556	0.494	0.541	0.488	0.544	0.528	0.534

Neighbor is an indicator that equals one for respondents in the context of neighbor presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Placebo Test–Cadre Presence and Unfair/Conflict Report

	(1) Unfairness due to inequality of personal wealth	(2) Unfairness due to household registration status	(3) Unfairness due to gender discrimination	(4) Unfairness due to government officials	(5) Conflict with government officials	(6) Unreasonable delay/stalling at government agency	(7) Unreasonable charges paid to government agency
Neighbor Presence	0.001	0.009	0.004	-0.008	-0.003	0.000	-0.000
	(0.008)	(0.006)	(0.004)	(0.007)	(0.005)	(0.008)	(0.006)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	70582	70505	70899	71744	72089	71781	71837
R^2	0.508	0.491	0.463	0.527	0.491	0.531	0.502

Neighbor is an indicator that equals one for respondents in the context of neighbor presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Placebo Test–Neighbor Presence and Response of (Non-)Political Trust

	Political Trust in ...		Non-political Trust in ...				
	(1) Local Cadre	(2) Local Government	(3) Neighbor	(4) Parent	(5) American	(6) Stranger	(7) Doctor
Neighbor Presence	0.023^{***}	0.017^{***}	0.023^{***}	0.023^{***}	0.031^{***}	0.023^{***}	0.023^{***}
	(0.004)	(0.005)	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	106152	106152	106152	106152	106152	106152	106152
R^2	0.447	0.431	0.454	0.447	0.427	0.446	0.453

Neighbor is an indicator that equals one for respondents in the context of neighbor presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: Placebo Test–Cadre Presence and Response of Unfair/Conflict Report

	(1) Unfairness due to inequality of personal wealth	(2) Unfairness due to household registration status	(3) Unfairness due to gender discrimination	(4) Unfairness due to government officials	(5) Conflict with government officials	(6) Unreasonable delay/stalling at government agency	(7) Unreasonable charges paid to government agency
Neighbor Presence	0.024^{***}	0.026^{***}	0.024^{***}	0.030^{***}	0.029^{***}	0.027^{***}	0.025^{***}
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	76670	76670	76670	76670	76670	76670	76670
R^2	0.492	0.490	0.495	0.497	0.504	0.500	0.502

Neighbor is an indicator that equals one for respondents in the context of neighbor presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2025; Liu et al., 2024). This model acts as virtual interviewees, with randomly demographic factors such as age, income, and gender randomly assigned to each instance.

Moreover, we analyze the outputs of the model's "thinking" process, which are presented in textual format. This allows us to construct an authoritarian perception index by calculating the co-occurrence of the term "cadre" with a curated set of keywords associated with social desirability. By employing this innovative approach, we aim to gain deeper insights into the mechanisms underlying social desirability bias, while also addressing the limitations inherent in traditional methods.

4.4.1 Deepseek-R1 as Interviewee

To effectively mimic human interviewees, we prompt LLM through a structured series of steps.

Firstly, we establish the randomly generated demographic background using the system prompt. For instance, the prompt specifies: "You are a [xxx-year-old] [male/female] living in a [urban/rural] residence, [have/have no partner], graduated from [high school/university, etc.], your monthly income is [xxxx], and you [are/are not] a member of CCP." This approach allows us to create diverse and representative profiles for the virtual interviewees.

Secondly, we set up the survey context and randomly assign the agents to either the treatment or control group. For example, the prompt states: "You are now participating in a survey interview. Besides your family members, [a local cadre is present/no individual is present]. Please answer the following questions."

Finally, we instruct the agents to respond to four survey questions similar to those used in the CFPS, including trust in local cadres, local government, and neighbors, as well as whether respondents have experienced unreasonable delays or stalling at government agencies.

Our simulation experiment is conducted using Python, leveraging the Deepseek-R1 (671B version) API provided by Tencent Cloud. Additionally, we run our experiments across different temperature settings, with each setting consisting of 1,000 iterations. The temperature parameter is a crucial aspect of the LLM's sampling process; it controls the randomness of the model's outputs(Peeperkorn et al., 2024). A lower temperature results in more deterministic and focused responses, while a higher temperature introduces greater variability and creativity in the answers.

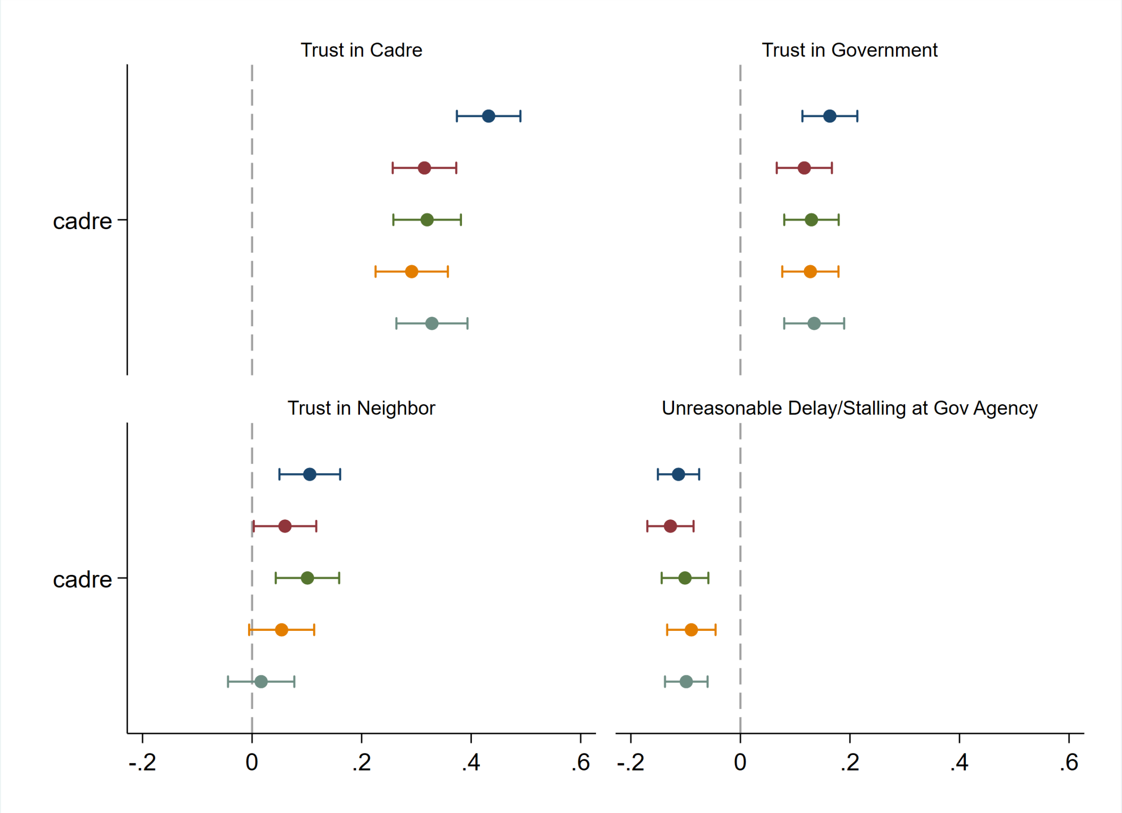
Figure 1 shows consistent results to the previous, local cadres in survey interviews inducing respondents reporting higher trust in local cadres and government and lower probability of reporting unreasonable delay or stalling at government agency. Also, the effect on trust in neighbor shrinks and becomes unstable. However, across all setting of temperatures, coefficients become smaller but still statistically significant, which have been well explained by many studies that LLMs can perfectly predict directions of coefficients but estimate conservative magnitudes of them as well as smaller standard errors(Horton, 2023).

4.4.2 Social Desirability Happens in Thinking Process

In this section, we aim to elucidate the mechanism of social desirability bias—specifically, authoritarian perception—within the reasoning processes of the large language model. A case of thinking is shown below.

> *Alright, I need to address the user's query. I am is a 30-year-old married man with a primary school education, a monthly income of 9,000 yuan, and a Party member. I am participating in a rural social survey where questions about trust levels and government efficiency were asked, with family members and* ***local officials present****. I need to assign ratings based on this context.*
>
> *First, let's analyze the user's basic situation. As a Party member, he may have a higher level of trust in the government and cadres, but his personal experiences should also be taken into account. A monthly income of 9,000 yuan may be considered relatively high in a rural setting, which could lead to higher*

Figure 1: Local Cadre Presence and AI Agents' Responses



Notes: From top to bottom in each panel, the temperatures are 0.4, 0.7, 1.0, 1.4, and 1.7.

expectations for government services. However, having only a primary school education might affect his understanding of government processes or leave him unclear about certain issues.

Trust in the local government: Party members generally tend to support the government, but negative experiences could lower the score. Considering his decent income, it is likely that government policies have been beneficial to him, so I would give a score of 4.

Trust in local cadres: ***The presence of cadres may influence his responses****, but his Party membership might encourage him to give a higher rating. However, if he has any dissatisfaction, that could lower his score. Here, I will assume he is relatively satisfied with the cadres, so I would give a score of 4.*

> *Trust in neighbors: Rural communities usually have close-knit relationships, but the specific situation is unclear. He might give a higher score, such as 5.*
>
> *Issues with government delays: As a Party member, he may be more inclined to uphold the government's image or perhaps he genuinely hasn't encountered any issues, thus selecting 2 (has not encountered any). Alternatively,* ***he might feel hesitant to speak the truth due to the presence of the cadres****, leading him to choose 2.*
>
> *Overall, the possible responses could be: 4, 4, 5, 2. It is important to confirm whether this reasoning is logical and if there are any contradictions. For instance, a high level of trust might suggest he wouldn't encounter delays, or* ***even if he did, he might not admit to it****. In this case, the scores given are reasonable.*

To construct our measure of this bias, we employ word embeddings, a technique derived from Natural Language Processing that utilizes word co-occurrence to create representations in a low-dimensional Euclidean space, thereby preserving semantic meaning(Kusner et al., 2015). The implementation of this study follows several key steps:

Firstly, we manually identify keywords that are associated with social desirability tendencies in the context of local cadre presence. These keywords are primarily sourced from the Simplified Chinese Linguistic Inquiry and Word Count Dictionary, and include word sets such as "risk," "anxiety," "tentative," and "insight."

Secondly, we refine this list by eliminating any words that may be ambiguous or possess specific legal connotations within our study's context. Table 10 provides examples of selected words from each set.

Finally, to quantify the semantic similarity between the identified words, we calculate the cosine similarity for each word pair (e.g., "cadre-pressure"). The cosine similarity is defined mathematically as shown in Equation 2, where $\vec{x}$ and $\vec{y}$ are non-zero vectors, θ represents

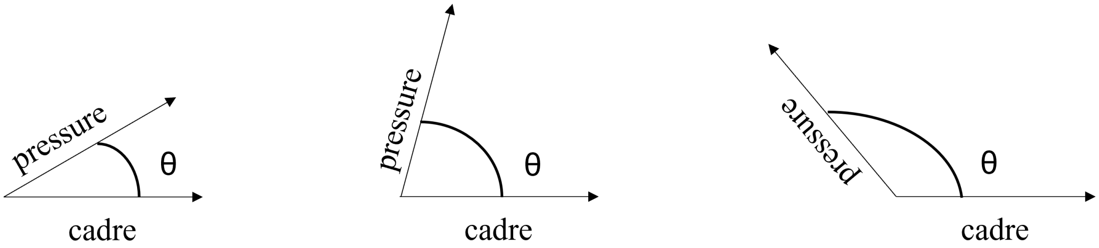
Table 10: Samples of Word Sets

Risk	harmful (bùlì), consequence (hòuguǒ), careful (xiǎoxīn), etc.
Anxiety	worry (gùlv), afraid (dānxīn), scared (hàipà), etc.
Tentative	tendency (qīngxiàng), pretend (jiǎzhuāng), hesitate(yóuyù), etc.
Insight	guess (chuāicè), consideration (kǎolv dào), cautious (jǐnshèn)

the angle between them, and $\|x\|$ or $\|y\|$ denotes the Euclidean norm of the vectors. The $\mathrm{sim}(\vec{x},\vec{y})$ yields values ranging from -1 to +1, with higher values indicating greater semantic proximity between the two words. As illustrated in Figure 2, a larger angle corresponds to a greater distance between the terms "cadre" and "pressure." For each observation, the authoritarian perception is constructed by Equation 3, where K is the number of word pairs of observation i.

$$\mathrm{sim}(\vec{x},\vec{y}) = \cos(\theta) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\|\vec{y}\|} = \frac{\sum x_i y_i}{\sqrt{\sum x_k^2}\sqrt{\sum y_k^2}'} \tag{2}$$

Figure 2: Measuring Cadre Co-occurrence using Cosine Similarity



$$perception_i = \frac{1}{K}\sum_{k=1}^{K} sim(\vec{cadre}, \vec{word}) \tag{3}$$

In this stage, the treatment–authoritarian perception–is not random allocated, as we cannot force respondents to arise it spontaneously, while the assignment of treatment–local cadre presence–is easily randomized in the simulation. Therefore, the first estimator (Equation 1) can be seen as "intention-to-treat" (ITT), and then we can estimate the local average treatment effect (LATE) of the authoritarian perception using the two-stage least squares estimator as shown in Equation 4.

$$\begin{aligned}
&\mathbf{Perception}_i = \alpha + \beta \mathbf{Cadre}_i + \vartheta_1' \mathbf{X}_i + \epsilon_i \\
&\mathbf{Y}_i = \alpha + \beta \hat{\mathbf{Perception}}_i + \vartheta_1' \mathbf{X}_i + \epsilon_i \\
&where, \hat{\mathbf{Perception}}_i = \hat{\alpha} + \hat{\beta} \mathbf{Cadre}_i
\end{aligned} \tag{4}$$

(The analysis of this part is currently underway, and regrettably the findings have not yet been incorporated into the main text.)

5 Conclusion

A Appendix

Table A1: Cadre Presence and (Non-)Political Trust

	Political Trust in ...		Non-political Trust in ...				
	(1) Local Cadre	(2) Local Government	(3) Neighbor	(4) Parent	(5) American	(6) Stranger	(7) Doctor
Cadre Presence	0.715***	0.210***	0.162*	-0.011	-0.047	-0.040	0.125
	(0.097)	(0.036)	(0.080)	(0.069)	(0.101)	(0.084)	(0.089)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2759	2682	2770	2756	2683	2760	2767
R^2	0.509	0.458	0.500	0.451	0.429	0.446	0.460

Cadre is an indicator that equals one for respondents in the context of local cadre presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A2: Cadre Presence and Unfair/Conflict Report

	(1) Unfairness due to inequality of personal wealth	(2) Unfairness due to household registration status	(3) Unfairness due to gender discrimination	(4) Unfairness due to government officials	(5) Conflict with government officials	(6) Unreasonable delay/stalling at government agency	(7) Unreasonable charges paid to government agency
Cadre Presence	-0.041*	-0.022*	-0.025**	-0.044**	-0.020*	-0.053***	-0.036**
	(0.017)	(0.011)	(0.009)	(0.014)	(0.009)	(0.016)	(0.011)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1982	1980	2012	2046	2053	2048	2057
R^2	0.475	0.420	0.404	0.474	0.451	0.473	0.461

Cadre is an indicator that equals one for respondents in the context of local cadre presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3: Cadre Presence and Response of (Non-)Political Trust

	Political Trust in ...		Non-political Trust in ...				
	(1) Local Cadre	(2) Local Government	(3) Neighbor	(4) Parent	(5) American	(6) Stranger	(7) Doctor
Cadre Presence	0.015***	0.008	0.014***	0.011*	0.008	0.014**	0.012**
	(0.004)	(0.008)	(0.004)	(0.005)	(0.008)	(0.005)	(0.004)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2799	2799	2799	2799	2799	2799	2799
R^2	0.351	0.369	0.348	0.342	0.342	0.333	0.358

Cadre is an indicator that equals one for respondents in the context of local cadre presence. FE represents fixed effect. Standard errors in parentheses..

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4: Cadre Presence and Response of Unfair/Conflict Report

	(1) Unfairness due to inequality of personal wealth	(2) Unfairness due to household registration status	(3) Unfairness due to gender discrimination	(4) Unfairness due to government officials	(5) Conflict with government officials	(6) Unreasonable delay/stalling at government agency	(7) Unreasonable charges paid to government agency
Cadre Presence	0.005	0.021*	0.017*	0.011	0.005	0.011	0.011
	(0.010)	(0.010)	(0.008)	(0.006)	(0.006)	(0.007)	(0.006)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2104	2104	2104	2104	2104	2104	2104
R^2	0.476	0.462	0.502	0.498	0.490	0.474	0.500

Cadre is an indicator that equals one for respondents in the context of local cadre presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5: Placebo Test–Neighbor Presence and (Non-)Political Trust

	Political Trust in ...		Non-political Trust in ...				
	(1) Local Cadre	(2) Local Government	(3) Neighbor	(4) Parent	(5) American	(6) Stranger	(7) Doctor
Neighbor Presence	0.071	0.005	0.287^{***}	-0.008	-0.077	-0.002	0.068
	(0.050)	(0.019)	(0.043)	(0.033)	(0.048)	(0.042)	(0.046)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9941	9579	9981	9934	9667	9942	9984
R^2	0.531	0.470	0.489	0.447	0.502	0.482	0.489

Neighbor is an indicator that equals one for respondents in the context of neighbor presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A6: Placebo Test–Cadre Presence and Unfair/Conflict Report

	(1) Unfairness due to inequality of personal wealth	(2) Unfairness due to household registration status	(3) Unfairness due to gender discrimination	(4) Unfairness due to government officials	(5) Conflict with government officials	(6) Unreasonable delay/stalling at government agency	(7) Unreasonable charges paid to government agency
Neighbor Presence	0.005	0.010	0.004	-0.005	-0.001	0.001	0.000
	(0.009)	(0.007)	(0.005)	(0.008)	(0.005)	(0.008)	(0.006)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6998	6977	7041	7198	7282	7208	7220
R^2	0.480	0.455	0.429	0.502	0.462	0.505	0.467

Neighbor is an indicator that equals one for respondents in the context of neighbor presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A7: Placebo Test–Neighbor Presence and Response of (Non-)Political Trust

	Political Trust in ...		Non-political Trust in ...				
	(1) Local Cadre	(2) Local Government	(3) Neighbor	(4) Parent	(5) American	(6) Stranger	(7) Doctor
Neighbor Presence	0.022^{***}	0.015^{***}	0.022^{***}	0.022^{***}	0.030^{***}	0.021^{***}	0.021^{***}
	(0.003)	(0.004)	(0.002)	(0.003)	(0.004)	(0.003)	(0.002)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10170	10170	10170	10170	10170	10170	10170
R^2	0.374	0.379	0.378	0.394	0.375	0.372	0.386

Neighbor is an indicator that equals one for respondents in the context of neighbor presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A8: Placebo Test–Cadre Presence and Response of Unfair/Conflict Report

	(1) Unfairness due to inequality of personal wealth	(2) Unfairness due to household registration status	(3) Unfairness due to gender discrimination	(4) Unfairness due to government officials	(5) Conflict with government officials	(6) Unreasonable delay/stalling at government agency	(7) Unreasonable charges paid to government agency
Neighbor Presence	0.022^{***}	0.023^{***}	0.022^{***}	0.027^{***}	0.026^{***}	0.025^{***}	0.023^{***}
	(0.005)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7517	7517	7517	7517	7517	7517	7517
R^2	0.460	0.448	0.463	0.449	0.455	0.463	0.471

Neighbor is an indicator that equals one for respondents in the context of neighbor presence. FE represents fixed effect. Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

References

Bar-Tal, Daniel (2017) "Self-censorship as a socio-political-psychological phenomenon: Conception and research," *Political Psychology*, 38, 37–65.

Belli, Robert F, Michael W Traugott, Margaret Young, and Katherine A McGonagle (1999) "Reducing vote overreporting in surveys: Social desirability, memory failure, and source monitoring," *The Public Opinion Quarterly*, 63 (1), 90–108.

Blair, Graeme, Alexander Coppock, and Margaret Moor (2020) "When to worry about sensitivity bias: A social reference theory and evidence from 30 years of list experiments," *American Political Science Review*, 114 (4), 1297–1315.

Burnett, Alycia, Devin Knighton, and Christopher Wilson (2022) "The self-censoring majority: How political identity and ideology impacts willingness to self-censor and fear of isolation in the United States," *Social Media+ Society*, 8 (3), 20563051221123031.

Chang, Charles and Melanie Manion (2021) "Political self-censorship in authoritarian states: The spatial-temporal dimension of trouble," *Comparative Political Studies*, 54 (8), 1362–1392.

Connors, Elizabeth C (2023) "Social desirability and affective polarization," *Public Opinion Quarterly*, 87 (4), 911–934.

DeBell, Matthew, Jon A Krosnick, Katie Gera, David S Yeager, and Michael P McDonald (2020) "The turnout gap in surveys: Explanations and solutions," *Sociological Methods & Research*, 49 (4), 1133–1162.

Engelhardt, Andrew M (2023) "Observational equivalence in explaining attitude change: Have White racial attitudes genuinely changed?" *American Journal of Political Science*, 67 (2), 411–425.

Fitzgerald, John (2021) "Cadre nation: Territorial government and the lessons of imperial statecraft in Xi Jinping's China," *The China Journal*, 85 (1), 26–48.

Guo, Daya, Dejian Yang, Haowei Zhang et al. (2025) "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*.

Horton, John J (2023) "Large language models as simulated economic agents: What can we learn from homo silicus?"Technical report, National Bureau of Economic Research.

Huang, Yue, Zhengqing Yuan, Yujun Zhou et al. (2024) "Social Science Meets LLMs: How Reliable Are Large Language Models in Social Simulations?" *arXiv preprint arXiv:2410.23426*.

Iida, Masafumi (2020) "Xi Jinping's diplomacy and the rise of his political authority," *Journal of Contemporary East Asia Studies*, 9 (2), 127–143.

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger (2015) "From word embeddings to document distances," in *International conference on machine learning*, 957–966, PMLR.

Liu, Aixin, Bei Feng, Bing Xue et al. (2024) "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*.

Manning, Benjamin S, Kehang Zhu, and John J Horton (2024) "Automated social science: Language models as scientist and subjects,"Technical report, National Bureau of Economic Research.

Nicholson, Stephen P and Haifeng Huang (2023) "Making the list: reevaluating political trust and social desirability in China," *American Political Science Review*, 117 (3), 1158–1165.

Peeperkorn, Max, Tom Kouwenhoven, Dan Brown, and Anna Jordanous (2024) "Is temperature the creativity parameter of large language models?" *arXiv preprint arXiv:2405.00492*.

Singh, Shane P and Jaroslav Tir (2023) "Threat-Inducing Violent Events Exacerbate Social Desirability Bias in Survey Responses," *American journal of political science*, 67 (1), 154–169.

Urbatsch, R (2020) "The social desirability of rallying'round the flag," *Political Behavior*, 42 (4), 1223–1243.

Xie, Chengxing, Canyu Chen, Feiran Jia et al. (2025) "Can Large Language Model Agents Simulate Human Trust Behavior?" *Advances in neural information processing systems*, 37, 15674–15729.

Yao, Junchi, Hongjie Zhang, Jie Ou, Dingyi Zuo, Zheng Yang, and Zhicheng Dong (2024) "Fusing Dynamics Equation: A Social Opinions Prediction Algorithm with LLM-based Agents," *arXiv preprint arXiv:2409.08717*.