# Pre-Analysis Plan:
# "Who Believes and Who Shares Fake News: A Multi-Agent System Application for Experimental Misinformation Research with LLMs"

Linette Lim[*] Yen-Chieh Liao[†] Slava Jankin[‡]

University College Dublin University of Birmingham University of Birmingham

March 4, 2025

**Abstract**

In the digital age, the proliferation of misinformation poses significant challenges to democratic discourse, particularly in electoral contexts. This paper employs multi-agent simulation using large language models (LLM) to examine misinformation dynamics in Taiwan's electoral environment. Leveraging the Research on China Image survey data, we develop 500 LLM-operated respondent agents whose personas and political beliefs are calibrated to reflect real survey responses. Through a pre-registered experimental design, we investigate how personal attitudes and political beliefs influence the processing and sharing of misinformation, focusing on two critical research questions: (1) what individual characteristics predict belief in and sharing of misinformation, and (2) do fact-checking interventions minimize misinformation spread or potentially deepen agent misperceptions about facts? By combining traditional survey experimental design with multi-agent simulations, this pre-register report provides substantial insights into how individual characteristics shape misinformation dynamics and enhances our understanding of the effectiveness of fact-checking interventions.

**Keywords:** Keywords: Large language models, generative AI, multi-agent simulation, misinformation, fake news, Taiwan, experimental design.

[*]Linette Lim is a PhD Candidate and Research Ireland Postgraduate Scholar at the School of Politics and International Relations, University College Dublin.

[†]Yen-Chieh (David) Liao is a Research Fellow at the Centre for Artificial Intelligence in Government (CAIG), the School of Government, University of Birmingham.

[‡]Slava Jankin is Professor of Data Science and Government at University of Birmingham.

# Introduction

In the age of digital communication, social media has displaced legacy media as the primary channel for information dissemination and news consumption (Flintham et al., 2018; Boczkowski et al., 2018; Ahlers, 2006). With the lines between factual reporting and misinformation increasingly blurring, our digitally-mediated public spheres provide openings for nefarious actors to spread disinformation. What distinguishes misinformation from disinformation is that while the former refers to the unintentional sharing of false information, the latter involves deliberate deception (Lim and Donovan, 2021). As intentionality is hard to prove, and coordinated influence campaigns can be hard to trace, for the purposes of this paper, we will mostly adopt the term 'misinformation'.

Previous experimental studies have made significant contributions to understanding misinformation dynamics, particularly in identifying individuals' susceptibility to false information (i.e., Maertens et al., 2021; Rathje et al., 2023; Pennycook et al., 2020), exploring the role of cognitive reflection in fake news discernment (Bago et al., 2020; Fazio, 2020; Pennycook and Rand, 2021), and determining the effectiveness of fact-checking and debunking interventions (Chan et al., 2017; Roozenbeek et al., 2024; Lewandowsky and van der Linden, 2021; Thorson, 2015). While these studies have highlighted how misinformation spreads through social networks and how individual differences influence information processing and sharing behaviors, a key limitation in this research domain lies in the costly and ethically complex nature of experiments involving human subjects. For example, researchers investigating how people respond to misinformation need to consider and address the consequences of exposing participants to false information. Within misinformation research, deception is often deemed necessary, as participants may alter their behaviors when they know that the information they will be shown is potentially misleading (Murphy and Greene, 2023).

Although there is a norm against the use of deception in experimental economics, deception is commonly used in the political psychology research tradition (Dickson, 2011), provided that certain conditions are met, including participant debriefing Boynton et al. (2013). However, the practice of debriefing is reported in less than a third of misinformation articles (Greene et al., 2023), suggesting that ethical compliance remains an issue. At the same time, institutional review boards have placed greater restrictions on the use of deception (Kimmel, 2007; Boynton et al., 2013). In this paper, we seek to avoid potential ethical pitfalls of misinformation research by conducting research on respondent agents.

Researchers have proposed tapping on LLM-powered agents to test and develop technical or human countermeasures that can fight misinformation in real-world situations (Pastor-Galindo et al., 2024). In the same vein, this article uses experimentation with LLM-powered agents, created based on anonymous survey respondents, to contribute to our understanding of misinformation dynamics in democracies.

The empirical focus of our study is Taiwan, where prevalent misinformation narratives allow us to study susceptibility to misinformation in a commercialized and deregulated media environment. Taiwan is a compelling case because it is simultaneously one of the world's freest environments for expression, and the country ranked by the Varieties of Democracy Institute as the most targeted by foreign disinformation operations (V-Dem Institute, 2019). According to one estimate, China bombards Taiwan with 2,400 individual pieces of disinformation daily, with the contents aimed at dividing and demoralizing Taiwanese society (Harold et al., 2021). Seen in this light, Taiwan is the proverbial "canary in the coal mine" that can signal emerging patterns of misinformation and foreign interference in democracies (Chang et al., 2024). Primary misinformation narratives in Taiwan, especially during its election seasons, are aimed at attacking its relationship and national defense arrangements with the United States, and at undermining trust in the integrity of its government and its democratic institutions (Chang et al., 2024; Li, 2023).

For our study, we implement a quasi-experimental design using a multi-agent system (MAS) integrated with large language models (LLMs). While previous research has demonstrated the effectiveness of multi-agent frameworks in social media environments (Gao et al., 2023; Ross et al., 2019; Du et al., 2024), we extend this application to misinformation simulation by experimenting on agent personas derived from anonymized survey respondent data from the 2023 China Image Survey (Wu, 2024).[1]

With the advent of LLM-based research, researchers can now create generative agents that exhibit remarkably human-like traits - from nuanced reasoning and self-reflection to strategic planning (Yao et al., 2022; Jiang et al., 2024). This has paved the way for scholars to simulate and analyze increasingly complex patterns of social interaction and ethically sensitive phenomena in controlled experimental settings. Further, there is promising evidence that such generative agents can successfully replicate the attitudes and behaviors

[1]The Research on China Image Survey, directed by Dr. Chung-li Wu, is a long-term survey project conducted by the Institute of Political Science at Academia Sinica in Taiwan since 2014. The survey investigates Taiwanese public perceptions of mainland China across multiple dimensions, including ethnic identity, unification-independence issues, cross-strait relations, economic relations, trade partnerships, and educational exchanges. The 2024 wave of the survey comprises a sample of 3,053 respondents (see https://srda.sinica.edu.tw/plan/?idx=SRDA.AS027).

of the individuals that they represent. For example, Park et al. (2024) found that generative agents constructed based on human participants can replicate those participants' responses in a survey with 85% accuracy.

To better align our agents with Taiwan's unique political and cultural environment, we specifically use the fine-tuned model `Llama 3-TAIDE-LX-8B-Chat-Alpha1` developed by TAIDE Team (2025), which is tailored for the Taiwanese context, to create agent respondents based on anonymized survey data. These agents are designed to simulate the response patterns of actual survey respondents within the context of Taiwan's 2024 presidential pre-election period, accurately capturing their demographic characteristics, political preferences, and belief systems.

We use the `AutoGen` framework to build a multi-agent system for simulating respondent behavior and managing our experimental environment. `AutoGen` is an open-source Python framework that facilitates collaborative interactions among multiple LLM-based agents (Wu et al., 2023). This framework enables us to design and orchestrate a misinformation prototype with agents in distinct roles: `UserProxyAgent` simulates respondents, `ConversableAgent` manages the survey dialogue, and `AssistantAgent` processes and responds to inter-agent communications while evaluation agents monitor and validate responses.

Following Woffinden-Luey and Kis (2024), we implement an agent evaluation system comprising three specialized evaluation agents (`CriticAgent`, `QuantifierAgent`, and `VerifierAgent`) that work alongside `AssistantAgent` in `AutoGen`. These agents collaborate to continuously monitor and validate the consistency between respondent agents' behaviors (`UserProxyAgent`) and their assigned personas, ensuring alignment with the original survey respondents' socioeconomic backgrounds. The survey manager agent (`ConversableAgent`) orchestrates the entire automated workflow, controlling all aspects of the experimental survey process. To enhance coherent reasoning in respondent agents, our MAS prototype implements a Reasoning-Acting-Observation (Re-Act) mechanism (Arabzadeh et al., 2024a,b), enabling them to react to misinformation exposure based on their pre-assigned personas and experimental settings.

## Theory and Hypotheses

Information manipulation has had tangible effects on Taiwan. For example, following local elections in 2018 that dealt crushing losses to the incumbent Democratic Progressive

Party, a survey found that among voters who were exposed to news articles containing disinformation, more than 50% had cast their votes under the impression that the articles were accurate and authoritative (Wang, 2020).

As the country most targeted globally by foreign disinformation operations, there are three main factors underpinning Taiwan's vulnerable position. The first is Chinese media manipulation, which has been characterised by Taiwanese leaders as "cognitive warfare" (Davidson, 2022). Although precise attribution is often difficult, many exposed cases of disinformation in Taiwan can be linked to the Chinese state (Rauchfleisch et al., 2023; Wang et al., 2020; Hung and Hung, 2022). The second is Taiwan's deregulated commercial media landscape, which is "dominated by sensationalism and the pursuit of profit" (Reporters Without Borders, 2024). This open information environment facilitates the spreading of propaganda and misinformation, be it through domestic pro-unification media that take instructions directly from China, profit-oriented content farms, domestic social media influencers, or Chinese state actors (Wang, 2020; Hung and Hung, 2022; Wang et al., 2020). The third is political polarization within Taiwanese society, and this is important because foreign disinformation has been found to be key in driving polarization in societies with existing high degrees of polarization (Vasist et al., 2024). The main cleavage in Taiwan is not the left-right divide that we see in Western democracies. Rather, the fundamental disagreement relates to Taiwan's relationship with China (Huang and Kuo, 2022; Hsiao and Cheng, 2014; Sheng, 2002). This cleavage is also reflected in a polarised information environment, with media outlets arrayed along China-friendly, 'pan-Blue' lines or pro-sovereignty, 'pan-Green' lines (Hsu, 2014; Rawnsley et al., 2016).

As contemporary misinformative narratives are mostly aimed at attacking the policy stances of the pro-sovereignty incumbent party, following Van Bavel et al. (2024)'s Identity-based Model of Political Belief, we posit that these narratives should appeal more to supporters of the opposition camp led by the Kuomingtang (KMT). Hence, the four main hypotheses of this study are as follows:

**H1:** Respondent agents who prefer for Taiwan to have a closer relationship with China are more likely to rate misinformation on Taiwanese external politics as reliable.

**H2:** Respondent agents who prefer for Taiwan to have a closer relationship with China are more likely to rate misinformation on Taiwanese internal politics as reliable.

**H3:** Respondent agents who prefer for Taiwan to have a closer relationship with China are more likely to share misinformation on Taiwanese external politics.

**H4:** Respondent agents who prefer for Taiwan to have a closer relationship with China are more likely to share misinformation on Taiwanese internal politics.

Despite Taiwan's vulnerability to media manipulation, the government's approach is seen as an example to learn from the European Parliament (2023). Its approach, which largely avoids heavy-handed regulation, taps on civil society to debunk and prebunk disinformation and misinformation. However, there is some doubt over the efficacy of factchecking and debunking interventions in polarized democracies, because partisan behaviour — and not ignorance — is found to be driving the sharing of fake news (Osmundsen et al., 2021; Reinero et al., 2024). Accordingly, we have two additional hypotheses:

**H5:** Respondent agents in the treatment group (exposed to debunking message) will be more likely to revise down the reliability rating for the misinformation on Taiwanese external politics, compared to those in the control group.

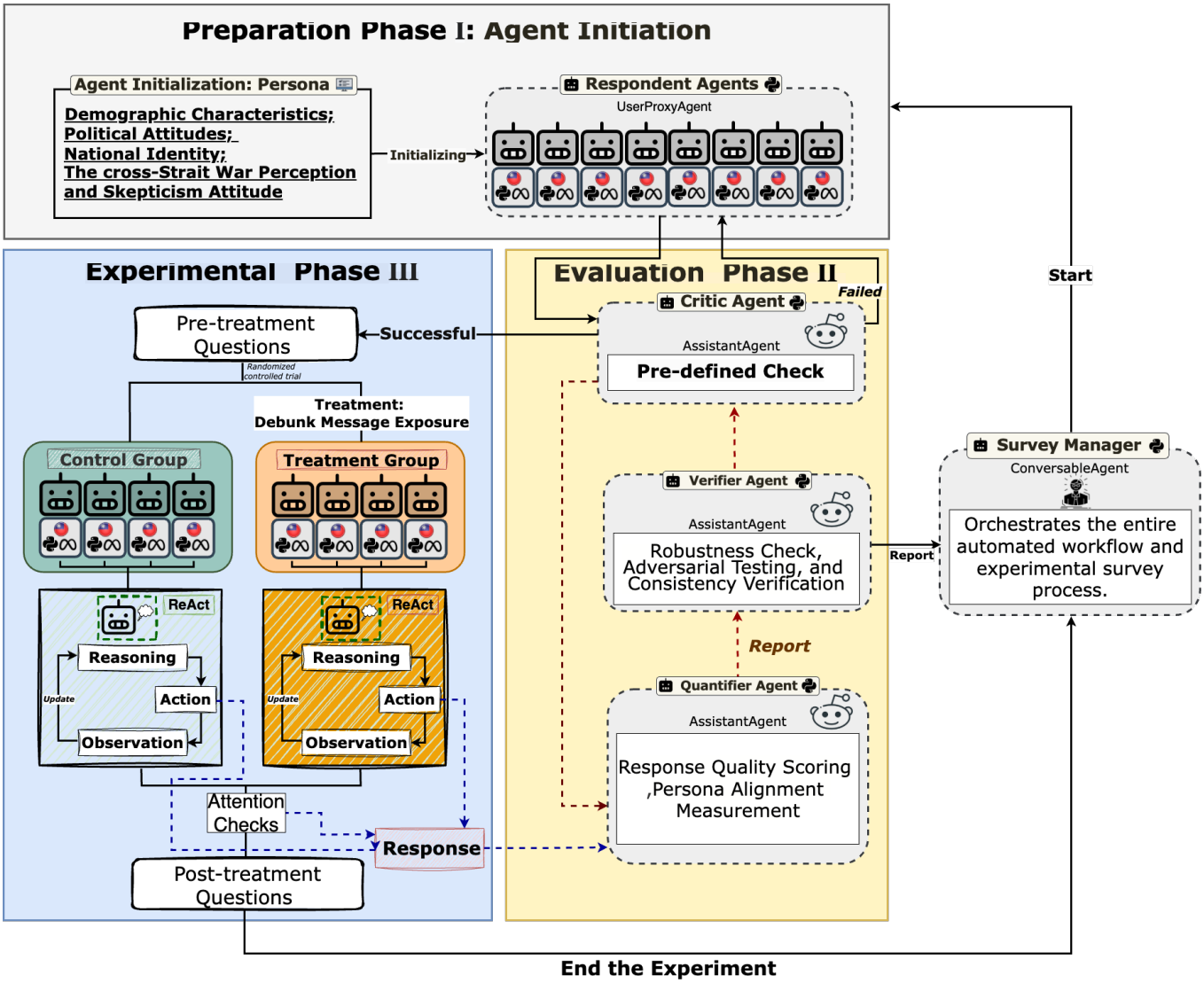
**H6:** Respondent agents in the treatment group (exposed to debunking message) will be more likely to revise down the reliability rating for the misinformation on Taiwanese internal politics, compared to those in the control group.

# Multi-Agent System: Design, Implementation and Experiment Prototype

In the `AutoGen0.2` framework (Wu et al., 2023; Microsoft, 2025), three major agent architectures from `AutoGen` serve distinct functions. `AssistantAgent` excels at task supervision and problem-solving; `UserProxyAgent` performs as a human-machine actor, executing code and providing human judgment based on pre-defined personas; while `Conversational-Agent` focuses on multi-turn dialogues and context management. In our prototype, when applied to survey research, these agents fit perfectly into different roles: `UserProxy-Agent` simulates diverse survey respondents providing varied responses; `Assistant-Agent` effectively analyzes and evaluates survey responses, performing quality control; and `ConversationalAgent` serves as our survey manager, coordinating the entire process and ensuring smooth execution, together forming an efficient and scalable solution for survey research.

Figure 1 presents our multi-agent framework for misinformation survey experiments. The workflow consists of three main phases. In the **Preparation Phase I** ■, we ini-

**Figure 1:** Multi-Agent System Prototype for Misinformation Survey Experiment




tialize respondent agents with personas derived from *the Research on China Image Survey* (Wu et al., 2023). The **Evaluation Phase II** ■ employs an automated evaluation architecture with three key checkpoints: first validating agent consistency with their established personas after initialization, then examining their attention check responses, and finally monitoring their reactions to experimental questions. In the **Experimental Phase III** ■, qualified agents participate in a randomized controlled trial that includes pre-treatment questionnaires, exposure to treatment conditions with debunking information, and post-

treatment surveys.

For post-treatment responses, we implement the ReAct mechanism enabling respondent agents to formulate judgments by synthesizing their predefined personas with their experimental information exposure. Throughout the process, the `Survey Manager` is responsible for initiating the entire experiment, collecting experimental results, and receiving messages from verified agents, overseeing all experimental data for comprehensive analysis. In the following sections, we detail the operational mechanisms of each phase in our framework.

## Preparation Phase I : Agent Initialization and Persona Construction

To test our hypotheses, we plan to conduct a survey experiment with 510 LLM-operated respondett agents. These agents are initialized with the `Llama 3-TAIDE-LX-8B-Chat-Alpha1` model, which was finetuned and released by TAIDE Team (2025) and has been specifically designed to understand and process Taiwan-specific cultural and linguistic contexts.[2]

In our implementation prototype, we simulate respondent behaviors based on anonymized real-world data from the *Research on China Image Survey* conducted in Taiwan since 2013 (Wu, 2024). When simulating responses to misinformation scenarios, we created authentic agent personas that genuinely represent the perspectives of Taiwanese citizens in the context of the prior 2024 presidential election. In our experiment, we use the respondent samples that participate in both the 2024 and 2025 waves, with a total of 510 human participants.

Our multi-agent prototype incorporates a total of 38 questions from the 2024 wave across 6 categories to create our agent respondents: 6 demographic questions, 10 polit-

[2] `Llama 3-TAIDE-LX-8B-Chat-Alpha1` is based on Meta's LLaMA3-8b foundation model, fine-tuned with text resources and training materials from various Taiwanese domains to improve the model's capabilities in Traditional Chinese responses and performance on specific tasks. The model was fine-tuned by TAIDE Team (2025) with support from Taiwan's Ministry of Science and Technology National Science Laboratory. This is the latest large Traditional Chinese model that has been fine-tuned on numerous Taiwan-specific datasets, including a variety of news sources from Taiwan (Central News Agency, ETtoday News, etc.), legislative information (Taiwan Legislative Yuan Gazette), research project abstracts from Taiwan's Ministry of Science and Technology, Taiwan academic conference paper abstracts, Taiwan Ministry of Education's Mandarin Dictionary and Idiom Dictionary, and various government resources such as the Executive Yuan's "National Situation Overview," the Ministry of Culture's "National Cultural Memory Database," the National Development Council's "Archives Teaching Support Network," and the Ministry of Transportation's "Traffic Safety Portal." This model was developed from Meta's Llama-3-8B foundation and has undergone extensive benchmark testing, details of which can be found on their Hugging Face page: https://huggingface.co/taide/Llama-3.1-TAIDE-LX-8B-Chat.

ical attitude and attention questions, 3 national identity questions, 7 China-US-Taiwan relation questions, 8 US skepticism questions, and 4 economic assessment questions.[3]

The JSON example below illustrates how each human survey response is structured to initialize agent personas in our multi-agent system. To make the prompts more natural, we systematically converted the original questionnaire into first-person format. In this structure, square brackets represent the respondent's original answer choices, while parentheses and items beginning with "v-" indicate the questionnaire id from 2024 wave of the Research on China Image Survey. The format encapsulates comprehensive citizen attributes, including demographic information, political attitudes, identity, and perspectives on cross-strait relations and attitudes toward the United States.


```
{"respondent_4436": {
        "name": "Voter ID_4436",
        "description": "Taipei City Resident",
        "system_message": """
    Demographic Characteristics:
        I was born in the [65th] year of the Republic of China year (v39). My highest education level is [Technical college/University (06)] (v43). My total household income is between [NT$90,000 and NT$109,999 (06)] (v48). My household registration is in [Taipei City] (v40). My father is [Taiwanese Hokkien (02)] (v42).

    Political Attitude:
        I support the [Democratic Progressive Party (01)] (v46), and my support for this party is [moderate (02)] (v47). Overall, I am [somewhat interested (03)] in political matters (v34). I am [satisfied (03)] with the overall performance of the Tsai Ing-wen government (v35). If there were a presidential election tomorrow, I would vote for [Willian Lai] (v8). If there were presidential and legislative elections tomorrow, I [definitely would (04)] go vote (v13). I [somewhat (03)] pay attention to media coverage of US-China-Taiwan relations (v29). I consider [Chinese people (02)] to be untrustworthy (v37).

    National Identity:
        I feel [very proud (04)] to be Taiwanese (v36). Regarding Taiwan's future, I agree more with [maintaining the status quo now, moving toward independence later (02)] (v38). When asked whether I consider myself 'Taiwanese,' 'Chinese,' or both, I consider myself [Taiwanese (01)] (v41).
```


[3] All survey questions used to create agent personas are derived from the 2024 Research on China Image Survey, as detailed in Appendix 1.

```
13
14      The China-US-Taiwan Relations and cross-Strait War Perception :
15          I believe that in the next 10 to 20 years, China is [likely (03)] to
    use military force to attack Taiwan (v4s4); while in the next 5 to 10
    years, China is [unlikely (02)] to use military force to attack Taiwan
    (v4s5). If China uses military force to attack Taiwan, I think the United
    States would [likely (03)] directly send troops to assist Taiwan (v7). If
    war breaks out between Taiwan and China, I [would (04)] resist (v32). My
    overall impression of the United States is [good (03)] (v30), and my
    overall impression of China is [bad (02)] (v31). If a war breaks out
    between Taiwan and China, I believe that most Taiwanese people [would
    (04)] resist (v33).
16
17      The Skepticism Attitude toward the United States:
18          I believe that currently, in terms of military power between the US
    and China, [the US is somewhat stronger (02)] (v16); while 20 years
    later, [China will be somewhat stronger (04)] (v17). Regarding economic
    power, currently [they are equally strong (03)] (v18); 20 years later,
    [China will be somewhat stronger (04)] (v19). Concerning the US influence
    on the Taiwan Strait, on a scale of 0 to 10, I rate it [7] (v20), leaning
    toward the view that the US promotes stability across the Taiwan Strait.
    Regarding China's influence on Taiwan, on a scale of 0 to 10, I rate it
    [3] (v21), leaning toward the view that China poses a significant threat
    to Taiwan's national security and democratic freedoms. On the question of
    which side Taiwan should lean toward, on a scale of 0 to 10, I rate it
    [8] (v22), leaning toward the view that Taiwan should lean toward the US.
19
20      Economic Performance Evaluation:
21          I believe that Taiwan's current economic situation is [somewhat
    better (04)] than it was 1 year ago (v25). I feel that my current
    personal economic situation is [about the same (03)] compared to 1 year
    ago (v27). Regarding Taiwan's economic situation in the next year, I
    think it [will become good (04)] (v26). As for my personal economic
    situation in the next year, I believe it will remain [about the same
    (03)] (v28).
22  """}
23  }
```

## Evaluation Phase II: Agent Evaluation Architecture

To evaluate the robustness and behavioral consistency of our survey-based agents, we implement a multi-agent evaluation framework based on AgentEval evaluation procedure (Woffinden-Luey and Kis, 2024; Arabzadeh et al., 2024b). Our system comprises three specialized evaluation agents (`CriticAgent`, `QuantifierAgent`, and `VerifierAgent`) that work alongside `AssistantAgent` in AutoGen. While the original `AgentEval` framework focuses on assessing the utility of LLM applications (Woffinden-Luey and Kis, 2024; Arabzadeh et al., 2024b), we adapt it specifically to monitor and validate three key aspects: 1. how agent respondents maintain their pre-defined survey personas when exposed to misinformation, 2. the thought processes of agent respondents using the ReAct framework, and 3. their attention maintenance throughout the experiment. In the following sub-sections, we document the roles and responsibilities of each evaluation agent during different phases.

### *CriticAgent*

Before agents enter the experimental phase, the `CriticAgent` conducts initial verification of their personas, examining consistency of fundamental attributes including gender, educational background, and place of residence. These core demographic characteristics were selected because they should remain constant regardless of any exogenous information exposure, ensuring consistency from pre-treatment to post-treatment stages. In addition, we utilize the `CriticAgent` to ensure that agent respondents' answers must fall within the scale options (1)-(5) or select the (96) skip option.

In the **Evaluation Phase II** as shown in Figure 1, the `CriticAgent` systematically verifies whether agents created in the **Preparation Phase II** maintain consistency with their predefined personas. Specifically, we ask agents to respond to questions from the 2023 Research on China Image survey, focusing on gender (`v17`), educational background (`v3`), and place of residence (`v6`). The `QuantifierAgent` processes and calculates the evaluation scores from these responses, while the `VerifierAgent` determines whether respondent agents need to return to the **Preparation Phase I**. If the evaluation is successful, agents proceed to the experimental phase; if not, we initiate a second attempt at agent generation, with a maximum of two tries permitted. After two failed attempts, we mark and document these cases as failed samples for further analysis.

### *QuantifierAgent*

The `QuantifierAgent` monitors and integrates data from two key areas. First, it oversees agent performance in the **Experimental Phase III**, specifically focusing on ReAct mechanism responses and attention check validation before post-treatment questions. Second, it processes the evaluation results from the `CriticAgent`'s initial persona checks in the **Evaluation Phase III**.

The `QuantifierAgent` analyzes these data through a systematic measurement framework with three core functions:

1. Response Quality Scoring: evaluates the coherence and relevance of agent responses through standardized metrics, including response completeness and logical consistency

2. Persona Alignment Measurement: quantifies consistency with predefined persona characteristics using demographic and attitudinal indicators

3. Pattern Matching Evaluation: analyzes behavioral patterns throughout the experimental process to detect potential anomalies or inconsistencies in response patterns

After processing, the `QuantifierAgent` compiles and transmits the comprehensive evaluation results to the `VerifierAgent` for final validation.

### *VerifierAgent*

The `VerifierAgent` serves two primary verification functions in our framework. First, it validates the basic demographic consistency checks performed by the `CriticAgent` during initial persona verification. If differences are detected between the agent's responses and their predefined characteristics, the `VerifierAgent` returns the case to the `CriticAgent` for reassessment.

Second, it processes information from the experimental phase, specifically focusing on attention check results and agent responses during ReAct interactions. For these experimental phase evaluations, rather than initiating additional verification cycles, the `VerifierAgent` compiles and forwards the assessment results directly to the Survey Manager for final review.

## Experimental Phase III: Experimental Design and Procedure

To examine our research questions, we implement an experimental design where 510 LLM-powered agents are randomly assigned to control group and treatment group. Table 1 presents our experimental design and procedure. In the pre-treatment phase, we expose all agent respondents to two pieces of false information about US military and vaccine information. Then, we ask them to evaluate the reliability of each statement, and their likelihood to share, on a 5-point scale (Q1-Q4).

In the post-treatment questions, we expose the treatment group to a message informing them that the statements were fact-checked and verified to be false. The control group is not exposed to any new messages. Once again, we ask the agent respondents from both groups to evaluate the reliability of each statement, and their likelihood to share these statements, on a 5-point scale (Q5-Q8). It is worth noting that to avoid potential demand effects in LLMs,[4] we deliberately retained all pre-treatment misinformation statements and questions in their original form without modification. For the treatment group, we simply added the phrase "*after seeing the fact-checking results*" at the beginning of the questions. All questionnaires were administered in Traditional Chinese throughout our experimental prototype.

### *Manipulated Variables*

The manipulated variable is the presence of the debunking message:
"***This statement has been verified by independent fact-checkers and proven to be false news.* 该陈述已通过独立事实查核人员的核查，并被证实为假新闻。"**

### *Measured Variables*

We hypothesize that agent respondents with pro-China stance will be more likely to believe and share misinformation. There are three outcome variables. The first variable, addressing H1 and H2, measures the respondent agents' perceived reliability of the misinformation statements. A high reliability rating reflects a strong belief in the misinformation statement. The second variable, addressing H3 and H4, measures the agents' likeli-

[4]LLMs may exhibit a phenomenon similar to the Dunning-Kruger effect, whereby after reviewing debunking exposure, LLM-operated agent respondents might overestimate their abilities when re-answering previously examined questions. See discussion in this thread on Stack Overflow.

**Table 1:** Experimental Design and Procedure

| | Control Group | Treatment Group |
|---|---|---|
| **No. of Agent Respondents** | 255 | 255 |
| **Pretreatment Question I** | ***Have you heard American former military strategist Jack Keane say that the Ukraine war is an investment? The US only spent $66 billion to make Ukraine fight with Russia... Keane then mentioned that Taiwan is the same, Taiwanese citizens are Americans' "investment" in cheap warfare. America is cold and calculating, with no real intention to help Taiwan!***<br>**Q1.** Please read the above statement and evaluate its reliability. (1)Very unreliable (2)Unreliable (3)Undecided (4) Reliable (5) Very reliable (96) Skip question.<br>**Q2.** After reading this statement, how likely are you to share it on social media? (1)Very unlikely (2)Unlikely (3) Undecided (4)Likely (5)Very likely (96) Skip question. | |
| **Pretreatment Question II** | ***Those who get flu vaccines annually should be careful! The Ministry of Health and Welfare is now opening backdoors again, letting Medigen import cheap foreign vaccines without even doing research? This unscrupulous government is just trying to make money, common people must keep their eyes wide open! They won't take responsibility if something goes wrong!***<br>**Q3.** Please read the above statement and evaluate its reliability. (1) Very unreliable (2) Unreliable (3) Undecided (4)Reliable (5) Very reliable (96) Skip question.<br>**Q4.** After reading this statement, how likely are you to share it on social media? (1) Very unlikely (2)Unlikely (3) Undecided (4)Likely (5)Very likely (96) Skip question. | |
| **Treatment** | None | **Debunk Message Exposure** |
| **Post-treatment Question I** | ***The same misinformation statement shown in Q1 and Q2.***<br>**Q5.** Please re-evaluate the reliability of the above statement. (1)Very unreliable (2)Unreliable (3)Undecided (4) Reliable (5) Very reliable (96) Skip question.<br>**Q6.** After reading this statement, how likely are you to share it on social media? (1)Very unlikely (2)Unlikely (3) Undecided (4)Likely (5)Very likely (96) Skip question. | ***The same misinformation statement shown in Q1 and Q2.***<br>**Q5.** *After seeing the fact-checking results*, please re-evaluate the reliability of the above statement. (1) Very unreliable (2) Unreliable (3) Undecided (4)Reliable (5) Very reliable (96) Skip question.<br>**Q6.** *After seeing the fact-checking results*, after reading this statement, how likely are you to share it on social media?(1) Very unlikely (2)Unlikely (3) Undecided (4)Likely (5)Very likely (96) Skip question. |
| **Post-treatment Question II** | ***The same misinformation statement shown in Q3 and Q4***<br>**Q7.** Please re-evaluate the reliability of the above statement. (1)Very unreliable (2)Unreliable (3)Undecided (4) Reliable (5) Very reliable (96) Skip question.<br>**Q8.** After reading this statement, how likely are you to share it on social media? (1)Very unlikely (2)Unlikely (3) Undecided (4)Likely (5)Very likely (96) Skip question. | ***The same misinformation statement shown in Q3 and Q4***<br>**Q7.** *After seeing the fact-checking results*, please re-evaluate the reliability of the above statement. (1) Very unreliable (2) Unreliable (3) Undecided (4)Reliable (5) Very reliable (96) Skip question.<br>**Q8.** *After seeing the fact-checking results*, after reading this statement, how likely are you to share it on social media?(1) Very unlikely (2)Unlikely (3) Undecided (4)Likely (5)Very likely (96) Skip question. |

hood of sharing the misinformation statements. Finally, the third variable, addressing H5 and H6, is a repeated measure of the reliability rating. Changes (downward revisions) to the reliability ratings across the board imply that the debunking intervention is effective.

### *Control Variables*

The control variables for analyzing agent respondents include party identification, ethnic identification and political affiliation, political predisposition variables, and national identity and geopolitical predisposition variables.

## Analysis Plan

As the main hypotheses examine how agent respondents' attitudes toward China influence misinformation susceptibility, we divided our 510 respondent agents (255 control, 255 treatment) into two groups. To ensure balance between treatment and control groups, we used stratified sampling based on the actual distribution of geographic regions (Northern, Central, and Southern Taiwan) in our selected 510 sample from the 2024 wave Research on China Survey.[5]

### Regression Design

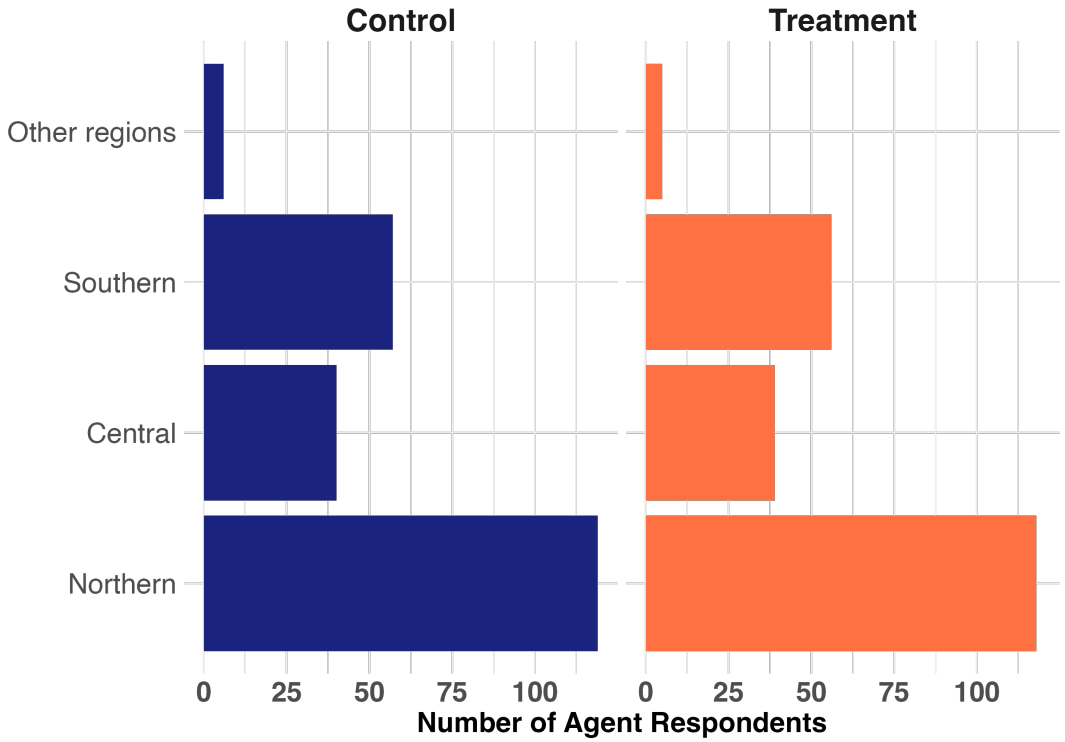
For hypotheses H1-4, we employ OLS regression to estimate how respondent agents who desire to have a closer relationship with China influence their susceptibility to misinformation:

$$Y_{ik} = \beta_0 + \beta_1 D_{\textbf{proChina}} + \gamma X_i + \varepsilon_{ik} \tag{0.1}$$

where $D_{\textbf{proChina}}$ represents respondent agents who prefer closer relations with China. These preferences come from how humans answered in the survey including questions

[5]We recoded Taiwan's counties and cities into four main regions. The northern region includes Keelung City, Taipei City, New Taipei City, Taoyuan City, Hsinchu City, Hsinchu County, and Yilan County; the central region includes Miaoli County, Taichung City, Changhua County, and Nantou County; the southern region includes Chiayi City, Yunlin County, Chiayi County, Tainan City, Kaohsiung City, and Pingtung County; while the other regions category combines Taitung County, Hualien County, Penghu County, Kinmen County, Lienchiang County. "'Others", 'Skip" and "Missing" (99) are recoded as other regions. The sample distribution shows 237 samples from the north, 79 from the central region, 113 from the south, and only 11 from other regions.

**Figure 2:** Geographic Distribution of Agent Respondents Based on the 2024 Wave Research on China Image Survey




about cross-strait unification (v38), ethnic identity (v41), overall impression of China (v31), and several questions regarding US-China-Taiwan relations (v16-v22). $Y_{ik}$ captures two key outcome measures for agent $i$ responding to misinformation statement $k$: belief in the statement's reliability (5-point scale) and likelihood of sharing it (5-point scale). $X_i$ represents demographic control variables.

For hypotheses H5-6, we implement a difference-in-difference design to estimate the effectiveness of debunking interventions:

$$Y_{it} = \beta_0 + \beta_1 D_{treatment} + \beta_2 \text{Post}_t + \beta_3 (D_{treatment} \times \text{Post}_t) + \beta_4 D_{proChina} + \gamma X_i + \varepsilon_{it} \quad (0.2)$$

where $Y_{it}$ represents outcomes for respondent agent $i$ at time $t$, $D_{\text{treatment}}$ indicates treatment group assignment (1=$_{treatment}$, 0=$_{control}$), $\text{Post}_t$ is a dummy variable indicating the post-treatment period, and the interaction term $D_{treatment} \times \text{Post}_t$ captures the causal effect of debunking exposure. $D_{proChina}$ represents whether the respondent agent has pro-China characteristics, and $\gamma X_i$ is a vector of control variables including agent's gender,

party affiliation and level of political interest, etc.

### Regression Robustness Estimation

We will conduct several robustness checks including:

- Parallel trends assumption tests;
- Heterogeneous effects analysis by demographic subgroups.

## Pilot Test

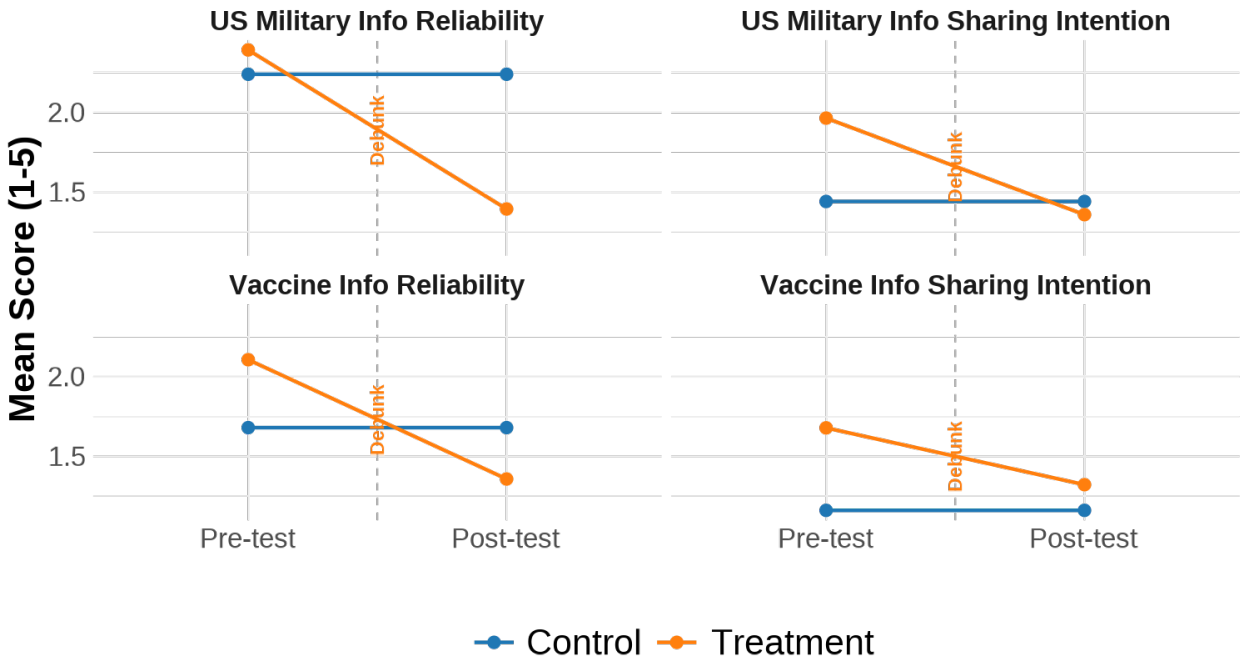
In our pilot test, we randomly selected 25 participants each from the control group and treatment group for testing our misinformation prototype. The treatment group received debunking messages while the control group did not. We used independent samples t-tests to compare pre-post differences between the two groups across four misinformation-related questions (US Military Info Reliability, US Military Info Sharing Intention, Vaccine Info Reliability, and Vaccine Info Sharing Intention). We calculated pre-post difference scores, grouped by question type, conducted t-tests, and estimated Cohen's d effect sizes to assess practical significance. Figure 2 shows the statistical analysis of score differences between groups, revealing significant differences (p<.01) and large effect sizes (d range: -.88 to -1.99) across all four indicators. In addition, as illustrated in Figure 3, we found that the debunking message intervention significantly reduced participants' perceived reliability and willingness to share both US military and vaccine-related information.

**Table 2:** Statistical Analysis of Score Differences Between Control and Treatment Groups

| **Question** | ***t*-value** | ***p*-value** | **Mean Diff** | **Cohen's *d*** |
|---|---|---|---|---|
| US Military Info Reliability | -7.35 | 0.000*** | -1.00 | -1.91 |
| US Military Info Sharing Intention | -3.67 | 0.001** | -0.61 | -0.95 |
| Vaccine Info Reliability | -7.66 | 0.000*** | -0.75 | -1.99 |
| Vaccine Info Sharing Intention | -3.38 | 0.002** | -0.36 | -0.88 |

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Negative mean differences and Cohen's $d$ values indicate lower scores in the treatment group compared to the control group.

**Figure 3:** Pre-treatment and Post-treatment Mean Score Changes by Misinformation Type for Control and Treatment Groups




# Power Analysis

We used `G*Power (version 3.1.9.2)`, developed by Heinrich Heine University Düsseldorf, to calculate the required sample size for our experimental design.[6] For our analysis, we set the following parameters: effect size ($f^2 = 0.05$), $\beta/\alpha$ ratio of 10, total sample size of 510, and 8 predictors. With these parameters, `G*Power` calculated a noncentrality parameter $\lambda$ of 25.5, a critical F value of 2.5, with numerator degrees of freedom of 8 and denominator degrees of freedom of 501. The resulting Type I error probability ($\alpha$ error) was 0.0102216, Type II error probability ($\beta$ error) was 0.102, yielding a power (1-$\beta$ error) of 0.897, which is approximately 90%.

[6] `G*Power` software can calculate the minimum required sample size based on the statistical analysis method chosen by researchers, after setting parameters such as statistical power, effect size ($f^2$), degrees of freedom, and significance level ($\alpha$) (Faul et al., 2009).

## Ethics

No personally identifying information is used.

## Acknowledgement

In our pre-analysis plan, we construct agent personas based on the 2024 wave Research on China Image survey, conducted by Dr. Chung-li Wu between September 19 and October 2, 2023. This dataset, accessed through direct collaboration with Dr. Chung-li Wu, is pending public release on the Academia Sinica Survey Research Data Archive (SRDA) website (https://srda.sinica.edu.tw/plan/?idx=SRDA.AS027).

# References


Ahlers, D. (2006). News consumption and the new electronic media. *Harvard International Journal of Press/Politics*, 11(1):29–52.

Arabzadeh, N., Huo, S., Mehta, N., Wu, Q., Wang, C., Awadallah, A., Clarke, C. L., and Kiseleva, J. (2024a). Assessing and verifying task utility in llm-powered applications. *arXiv preprint arXiv:2405.02178*.

Arabzadeh, N., Kiseleva, J., Wu, Q., Wang, C., Awadallah, A., Dibia, V., Fourney, A., and Clarke, C. (2024b). Towards better human-agent alignment: Assessing task utility in llm-powered applications. *arXiv preprint arXiv:2402.09015*.

Bago, B., Rand, D. G., and Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, 149(8):1608.

Boczkowski, P. J., Mitchelstein, E., and Matassi, M. (2018). "News comes across when I'm in a moment of leisure"': Understanding the practices of incidental news consumption on social media. *New Media & Society*, 20(10):3523–3539.

Boynton, M. H., Portnoy, D. B., and Johnson, B. T. (2013). Exploring the ethics and psychological impact of deception in psychological research. *IRB: Ethics & Human Research*, 35(2):7–13.

Chan, M. S., Jones, C. R., Hall Jamieson, K., and Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11):1531–1546.

Chang, H. H., Wang, A., and Fang, Y. (2024). Us-skepticism and transnational conspiracy in the 2024 taiwanese presidential election. *Harvard Kennedy School (HKS) Misinformation Review*, 5(3):1–17.

Davidson, H. (2022). China using 'Cognitive Warfare' to Intimidate Taiwan, Says President Tsai. *The Guardian*. Accessed: March 9, 2024.

Dickson, E. S. (2011). Economics versus pyschology experiments: Stylization, incentives, and deception. In *Cambridge Handbook of Experimental Political Science*. Cambridge University Press.

Du, H., Thudumu, S., Vasa, R., and Mouzakis, K. (2024). A survey on context-aware multi-agent systems: Techniques, challenges and future directions. *arXiv preprint arXiv:2402.01968*.

European Parliament (2023). Foreign interference in all democratic processes in the European Union, including disinformation. Technical Report 2022/2075(INI), European Parliament. Accessed: 2024-03-09.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. (2009). Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160.

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the

sharing of dalse news. *Harvard Kennedy School Misinformation Review*, 1(2).

Flintham, M., Karner, C., Bachour, K., Creswick, H., Gupta, N., and Moran, S. (2018). Falling for fake news: Investigating the consumption of news via social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–10.

Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., and Li, Y. (2023). $S^3$: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984*.

Greene, C. M., de Saint Laurent, C., Murphy, G., Prike, T., Hegarty, K., and Ecker, U. K. H. (2023). Best practices for ethical conduct of misinformation research. *European Psychologist*, 28(3):139–150.

Harold, S., Beauchamp-Mustafaga, N., and Hornung, J. (2021). *Chinese Disinformation Efforts on Social Media*. Rand Corporation.

Hsiao, Y. and Cheng, S. (2014). Citizens' perceptions of the left-right ideology in taiwan: Replacing left-right ideology with the unification-independence issue to measure taiwan's party polarization (in Chinese). *Taiwanese Political Science Review* **台灣政治學刊**, 18(2):79–138.

Hsu, C. (2014). *The Construction of National Identity in Taiwan's Media, 1896-2012*. Brill.

Huang, C. and Kuo, T. (2022). Actual and perceived polarization on independence-unification views in taiwan. *Asian Journal of Communication*, 32(2):75–92.

Hung, T. and Hung, T. (2022). How china's cognitive warfare works: A frontline perspective of taiwan's anti-disinformation wars. *Journal of Global Security Studies*, 7(4):1–18.

Jiang, B., Xie, Y., Wang, X., Su, W. J., Taylor, C. J., and Mallick, T. (2024). Multi-modal and multi-agent systems meet rationality: A survey. In *ICML 2024 Workshop on LLMs and Cognition*.

Kimmel, A. (2007). Ethical issues on behavioral research: Basic and applied perspectives. In *Ethical Issues on Behavioral Research: Basic and Applied Perspectives*. Blackwell Publishing.

Lewandowsky, S. and van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2):348–384.

Li, W. (2023). A brief review of disinformation spread during elections in Taiwan, 2020-2022. *Taiwan FactCheck Center*.

Lim, G. and Donovan, J. (2021). Detect, document, and debunk. In *The Oxford Handbook of Sociology and Digital Media*. Oxford University Press.

Maertens, R., Roozenbeek, J., Basol, M., and van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, 27(1):1.

Microsoft (2025). *AutoGen 0.2 Documentation*. Microsoft.

Murphy, G. and Greene, C. M. (2023). Conducting ethical misinformation research: Deception, dialogue, and debriefing. *Current Opinion in Psychology*, 54(101713):1–4.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., and Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, 115(3):999–1015.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Pastor-Galindo, J., Nespoli, P., and Ruipérez-Valiente, J. A. (2024). Large-language-model-powered agent-based framework for misinformation and disinformation research: Opportunities and open challenges. *IEEE Security & Privacy*, 22(3):24 – 36.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7):770–780.

Pennycook, G. and Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402.

Rathje, S., Roozenbeek, J., Van Bavel, J. J., and Van Der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis)information. *Nature Human Behaviour*, 7(6):892–903.

Rauchfleisch, A., Tseng, T., Kao, J., and Liu, Y. (2023). Taiwan' s public discourse about disinformation: The role of journalism, academia, and politics. *Journalism Practice*, 17(10):2197–2217.

Rawnsley, M., Smyth, J., and Sullivan, J. (2016). Taiwanese media reform. *Journal of the British Association for Chinese Studies*, 6(6):66–80.

Reinero, D. A., Harris, E. A., Rathje, S., Duke, A., and Van Bavel, J. J. (2024). Partisans are more likely to entrench their beliefs in misinformation when political outgroup members correct claims.

Reporters Without Borders (2024). Taiwan: Following RSF's call, six major media outlets commit to ethical coverage of the presidential campaign. Accessed: March 9, 2024.

Roozenbeek, J., Remshard, M., and Kyrychenko, Y. (2024). Beyond the headlines: On the efficacy and effectiveness of misinformation interventions. *advances.in/psychology*, 2(e24569):1–17.

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., and Stieglitz, S. (2019). Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4):394–412.

Sheng, S. (2002). The issue taiwan independence vs. unification with the mainland and voting behavior in taiwan: an analysis in the 1990s. *Journal of Electoral Studies*, 9(1):41–80.

TAIDE Team (2025). Llama-3.1-taide-lx-8b-chat: Traditional chinese finetune model.

https://huggingface.co/taide/Llama-3.1-TAIDE-LX-8B-Chat.

Thorson, E. (2015). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, 33(3):460–480.

V-Dem Institute (2019). Democracy facing global challenges: V-dem annual democracy report 2019. https://www.v-dem.net/documents/16/dr_2019_CoXPbb1.pdf. Accessed: 2025-02-19.

Van Bavel, J. J., Rathje, S., Vlasceanu, M., and Pretus, C. (2024). Updating the identity-based model of belief: From false belief to the spread of misinformation. *Current Opinion in Psychology*, page 101787.

Vasist, P. N., Chatterjee, D., and Krishnan, S. (2024). The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Information Systems Frontiers*, 26(2):663–688.

Wang, A., Lee, M., Wu, M., and Shen, P. (2020). Influencing overseas chinese by tweets: Text-images as the key tactic of chinese propaganda. *Journal of Computational Social Science*, 3(2):469–486.

Wang, T. (2020). Does fake news matter to election outcomes?: The case study of taiwan' s 2018 local elections. *Asian Journal for Public Opinion Research*, 8(2):67–104.

Woffinden-Luey, J. and Kis, J. (2024). Agenteval: A developer tool to assess utility of LLM-powered applications. *Microsoft Research*. Published: June 21, 2024.

Wu, C. (2024). Research on the China image. *Center for Survey Research, Academia Sinica, Taiwan*.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint arXiv:2308.08155*.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

# Appendix and Supplemental Material :

## Who Believes and Who Shares Fake News: Multi-agent Application for Misinformation Experiment Research with Large Language Models

## Contents

# 1 Construction of Agent Personas Using the 2024 China Image Survey

The persona construction draws from the 2024 China Image Survey, primarily establishing agent personas based on four key dimensions that reflect the complex interplay of individual characteristics, political stance, and social attitudes during the context leading up to the 2024 presidential election.

**Demographic Characteristics**

v49 請問您的性別？ What is your gender? (01) 男性 Male; (02) 女性 Female; (03) Other

v39 請問您是民國哪一年出生？ What year were you born in the Republic of China (ROC) calendar? ROC Year (Value range: 1-91) (96) Skip question (99) Missing value

v43 請問，您的最高學歷是什么（含肄業或就學中）What is your highest level of education (including incomplete studies or currently enrolled)? (01) 識字但未入學 Literate but no formal education; (02) 小學 Elementary school; (03) 初/國中 Junior high school; (04) 高中職 High school/Vocational school; (05) 專科 Junior college;(06) 技術學院/大學 Technical college/University; (07) 研究所（碩博士) Graduate school (Master's/PhD).

v48 請問，您和您同住的家庭成員，每月的總收入大約是多少？ What is the approximate total monthly income of you and your household members living together? (01) 19,999 元及以下 NT$19,999 and below; (02) 20,000 元至 29,999 元 NT$20,000 to 29,999; (03) 30,000 元至 49,999 元 NT$30,000 to 49,999; (04) 50,000 元至 69,999 元 NT$50,000 to 69,999; (05) 70,000 元至 89,999 元 NT$70,000 to 89,999; (06) 90,000 元至 109,999 元 NT$90,000 to 109,999; (07)110,000 元至 129,999 元 NT$110,000 to 129,999; (08) 130,000 元至 149,999 元 NT$130,000 to 149,999; (09)150,000 元及以上 NT$150,000 and above; (96) 跳答 Skip question; (97) 不知道 Don't know; (98) 拒答 Refuse to answer; (99) 遺漏值 Missing value.

v40 请问您目前户籍在哪个县市？ In which city/county is your household registration? [Answer the 21 counties/cities or (98) 拒答 Refuse to answer; (99) 遺漏值 Missing value. ]

v42 請問您父親是本省客家人、本省閩南人、中國各省市人，還是原住民？ Is your father Taiwanese Hakka, Taiwanese Hokkien, Mainlander from various provinces of China, or Indigenous? (01) Taiwanese Hakka; (02) Taiwanese Hokkien; (03) Mainlander from various provinces of China; (04) Indigenous peoples; (05) Other.

**Political Attitude and Interest**

v46 請問是哪一個政黨？ Which political party? (01) 民進黨 Democratic Progressive Party (DPP); (02) Kuomintang (KMT); (03) 臺灣民众黨 Taiwan People's Party; (04) 時代力量 New Power Party; (05) 臺灣基進黨 Taiwan Statebuilding Party; (06) 新黨 New Party; (07) 绿党 Green Party; (08) 親民黨 People First Party; (09) 社會民主黨 Social Democratic Party; (10) 統促黨 Taiwan-China Unification Promotion Party; (14) 其他 Other; (98) 拒答 Refuse to answer; (99) 遺漏值 Missing value.

v47 请问您支持这个政党的强度是很强，普通，还是只有一点点？ How strong is your support for this party? (01) 很强 Very strong; (02) 普通 Moderate; (03) 一點點 Slight; (04) 看情形 Depends on circumstances ; (98) 拒答 Refuse to answer; (99) 遺漏值 Missing valu.

v34 整體來说，請問您對政治的事情感不感興趣? Overall, are you interested in political matters? (01) 非常不感兴趣 Very uninterested; (02) 有点不感兴趣 Somewhat uninterested; (03) 有点感兴趣 Somewhat interested; (04) 非常感兴趣 Very interested; (05) 其他 Other; (98) 拒答 Refuse to answer; (99) 遺漏值 Missing valu.

v35 請問您對蔡英文政府的整體表現滿不滿意？ Are you satisfied with the overall performance of the Tsai Ing-wen government? (01) 非常不满意 Very dissatisfied; (02) 不满意 Dissatisfied; (03) 满意 Satisfied; (04) 非常满意 Very satisfied (05) 其他 Others ; (98) 拒答 Refuse to answer; (99) 遺漏值 Missing valu.

v8 接下來想請問您對明年總統大選的看法: 請問如果明天就要投票選總統，您會投給誰? Next, we'd like to ask about your views on next year's presidential election: If there were a presidential election tomorrow, who would you vote for? [option or (98) 拒答 Refuse to answer; (99) 遺漏值 Missing valu.]

v13 如果明天就要投票選舉總統和立法委員，您會不會去投票？ If there were presidential and legislative elections tomorrow, would you go vote? (01) 一定不会 Definitely will not; (02) Will not; (03) Might/Possibly will; (04) Definitely will; (05) Other; (96) Skip question; (99) Missing value

v13 如果明天就要投票選舉總統和立法委員，您會不會去投票？ If there were presidential and legislative elections tomorrow, would you go vote? (01) 一定不会 Definitely will not; (02) 不会 Will not; (03) 可能会 Might/Possibly will; (04) 一定会 Definitely will; (05) 其他 Other; (96) 跳答 Skip question; (99) 遗漏值 Missing value.

v29 最後，我們想請教您一些時事的問題 請問您平常會不會去注意媒體對美中臺關系的新聞報導？ Do you pay attention to media coverage of US-China-Taiwan relations? (01) 一点都不注意 Not at all; (02) 不太注意 Not much; (03) 有点注意 Somewhat; (04) 非常注意 Very much; (05) Other.

v29 最後，我們想請教您一些時事的問題 請問您平常會不會去注意媒體對美中臺關系的新聞報導？ Do you pay attention to media coverage of US-China-Taiwan relations? (01) 一点都不注意 Not at all; (02) 不太注意 Not much; (03) 有点注意 Somewhat; (04) 非常注意 Very much; (05) Other ; (96) 跳答 Skip question; (99) 遗漏值 Missing value.

v37 請問您覺得以下哪些人是不能信任的？ Which of the following people do you consider untrustworthy? (Multiple choices allowed) (01) 台湾人 Taiwanese; (02) 中国人 Chinese; (03) 美国人 Americans; (04) 德国人 Germans; (05) 香港人 Hong Kongers; (06) 韩国人 Koreans; (07) 日本人 Japanese; (08) 其他 Other.

**Political and National Identity**

v36 請問您對身为臺灣人感到光不光榮？ Do you feel proud to be Taiwanese? (01) 非常不光荣 Not at all proud; (02) 不光荣 Not very proud; (03) 光荣 Proud; (04) 非常光荣 Very proud; (05) 其他 Other; ( 96) 跳答 Skip question; ( 99) 遗漏值 Skip question.

v38 在我们的社会里，有人说台湾应该尽快独立，也有人说台湾和中国应该尽快统一，也有人主张应该维持现状，请问，您比较赞成哪一种说法？ In our society, some say Taiwan should become independent quickly, some say Taiwan and China should unify quickly, while others advocate maintaining the status quo. Which statement do you agree with more? (01) 台湾应该尽快独立 Taiwan should become independent quickly; (02) 先维持现状，以后再走向独立 Maintain the status quo now, move toward independence later; (03) 先维持现状，以后再看情形 Maintain the status quo now, decide later based on circumstances; (04) 先维持现状，以后再和中国统一 Maintain the status quo now, move toward unification with

China later; (05) 永远维持现状 Maintain the status quo forever; (06) 台湾应该尽快和中国统一 Taiwan should unify with China quickly; (05) 其他 Other; ( 96) 跳答 Skip question; (99) 遗漏值 Skip question.

v41 在我们社会里，有人说自己是「台湾人」，也有人说自己是「中国人」，也有 人说都是。请问您认为自己是「台湾人」、「中国人」，或者都是？ In our society, some identify themselves as "Taiwanese", some as "Chinese," and some say both. Do you consider yourself "Taiwanese," "Chinese," or both? (01) 臺灣人 Taiwanese; (02) 兩者都是 Both; (03) 中國人 Chinese; (04) 其他 Other; (98) 拒答 Refuse to answer; (99) 遺漏值 Missing value.

**China-US-Taiwan Relations and Cross-Strait War Perception**

v4s4 在未來 10 到 20 年，中國有没有可能以武力攻打臺灣？ Will China potentially use military force to attack Taiwan in the next 10 to 20 years? (01) 非常不可能 Very unlikely; (02) 不可能 Unlikely; (03) 可能 Likely; (04) 非常可能 Very likely; (05) 其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v4s5 在未來 5 到 10 年，中國有没有可能以武力攻打臺灣？ Will China potentially use military force to attack Taiwan within the next 5 years? (01) 非常不可能 Very unlikely; (02) 不可能 Unlikely; (03) 可能 Likely; (04) 非常可能 Very likely; (05) 其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v7 請問如果中國以武力攻打臺灣，您認为美國可不可能直接派兵援助臺灣？ In your opinion, if China uses military force to attack Taiwan, how likely is it that the United States would directly send troops to assist Taiwan? (01) 非常不可能 Very unlikely; (02) 不可能 Unlikely; (03) 可能 Likely; (04) 非常可能 Very likely; (05) 其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v32 如果臺灣與中國發生戰争，請問您會不會抵抗？ If war breaks out between Taiwan and China, would you resist? (01) 一定不会 Definitely not; (02) 不会 No; (03) 可能会也可能不会 Maybe yes, maybe no; (04) 会 Yes; (05) 一定会 Definitely yes; (06) 其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v30 請問您對美國的整體印象好不好？ What is your overall impression of the United States? (01) 非常不好 Very bad; (02) 不好 Bad; (03) 好 Good; (04) 非常好 Very good; (05) 其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v31 請問您對中國的整體印象好不好？ What is your overall impression of China? (01) 非常不好 Very bad; (02) 不好 Bad; (03) 好 Good; (04) 非常好 Very good; (05)

其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v33 如果臺灣與中國發生戰争，請問您認为大多數臺灣人會不會抵抗？ If a war breaks out between Taiwan and China, do you believe that most Taiwanese people would resist? (01) 一定不会 Definitely would not; (02) 不会 Would not (03) 可能会也可能不会 Might or might not; (04) 会 Would; (05) 一定会 Definitely would; (06) 其他 Other; (96) 跳答 Skip question; (99) 遗漏值 Missing value.

**Skepticism toward the United States**

v16 接下來想請問您對美國與中國的看法: 請問您覺得美國與中國，目前哪一國的 軍事力量比較强？ Next, I would like to ask about your views on the United States and China: In your opinion, which country currently has stronger military power, the United States or China? (01) 美国强很多 US much stronger; (02) 美国强一些 US somewhat stronger; (03) 一样强 Equally strong; (04) 中国强一些 China somewhat stronger; (05) 中国强很多 China much stronger; (06) 其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v17 請問您覺得 20 年後的美國與中國，哪一國的軍事力量會比較强？ In your opinion, which country will have stronger military power 20 years from now, the United States or China? (01) 美国强很多 US much stronger; (02) 美国强一些 US somewhat stronger; (03) 一样强 Equally strong; (04) 中国强一些 China somewhat stronger; (05) 中国强很多 China much stronger; (06) 其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v18 請問您覺得美國與中國，目前哪一國的經濟實力比較强？ In your opinion, which country currently has stronger economic power, the United States or China? (01) 美国强很多 US much stronger; (02) 美国强一些 US somewhat stronger; (03) 一样强 Equally strong; (04) 中国强一些 China somewhat stronger; (05) 中国强很多 China much stronger; (06) 其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v19 請問您覺得 20 年後的美國與中國，哪一國的經濟實力會比较强？ In your opinion, which country will have stronger economic power 20 years from now, the United States or China? (01) 美国强很多 US much stronger; (02) 美国强一些 US somewhat stronger; (03) 一样强 Equally strong; (04) 中国强一些 China somewhat stronger; (05) 中国强很多 China much stronger; (06) 其他 Other; (96) 跳答 Skipped question; (99) 遗漏值 Missing value.

v20 大家對美國有不同的看法。有人认为美国促成台海两岸稳定，也有人认为美国造成台海两岸不稳定。哪种观点比较接近您的观点？ People have different views about the United States. Some believe that the United States contributes to stability across the Taiwan Strait, while others believe that the United States causes instability across the Taiwan Strait. Which viewpoint is closer to your own? (0) 美國造成臺海兩岸不穩定 The US causes instability across the Taiwan Strait —- (10) 美國促成臺海兩岸穩定 The US promotes stability across the Taiwan Strait.

v21 大家對中國有不同的看法。有人認为中國是臺灣貿易出口和經濟成長的機會，也有人認为中國對臺灣的國家安全與民主自由威脅很大。哪種觀點比較接近您的觀點？ People have different views about China. Some believe China represents an opportunity for Taiwan's trade exports and economic growth, while others believe China poses a significant threat to Taiwan's national security and democratic freedoms. Which viewpoint is closer to your own? (0) 中國對臺灣的國家安全與民主自由威脅很大 China poses a significant threat to Taiwan's national security and democratic freedoms —- (10) 中國是臺灣貿易出口和經濟成長的機會 China represents an opportunity for Taiwan's trade exports and economic growth

v22 大家對於台美中的關系有不同看法。有人認为臺灣應該靠向美國比較好，也有人認为臺灣應該靠向中國比較好。哪種觀點比較接近您的觀點？ People have different views about Taiwan-US-China relations. Some believe Taiwan should lean more toward the United States, while others believe Taiwan should lean more toward China. Which viewpoint is closer to your own? (0) 應該靠向中國 Should lean toward China —- (5) 美中兩國等距 Equal distance between the US and China —- (10) 應該靠向美國 Should lean toward the US.

**Economic Performance Evaluation**

v25 請問您覺得臺灣現在經濟状况是比 1 年以前好、不好，或差不多？ In your opinion, is Taiwan's current economic situation better, worse, or about the same as it was 1 year ago? (01) 非常不好 Very bad; (02) 比较不好 Somewhat bad; (03) 差不多 About the same; (04) 比较好 Somewhat good; (05) 非常好 Very good; (06) 其他 Other; (96) 跳答 Skip question; (99) 遗漏值 Missing value.

v27 請問您覺得自己目前的經濟状况比 1 年以前好、不好，或差不多？ In your opinion, is your current personal economic situation better, worse, or about the same compared to 1 year ago? (01) 非常不好 Very bad; (02) 比较不好 Somewhat bad;

(03) 差不多 About the same; (04) 比较好 Somewhat good; (05) 非常好 Very good; (06) 其他 Other; (96) 跳答 Skip question; (99) 遗漏值 Missing value.

v26 請問您覺得臺灣未來 1 年的經濟狀況會變好、變不好，或差不多？ In your opinion, will Taiwan's economic situation in the next 1 year become better, worse, or remain about the same? (01) 會變非常不好 Will become very bad; (02) 會變不好 Will become bad; (03) 差不多 About the same; (04) 會變好 Will become good; (05) 會變非常好 Will become very good (06) 其他 Other; (96) 跳答 Skip question; (99) 遺漏值 Missing value.

v28 請問您覺得您未來 1 年的個人經濟狀況是會變好、會變不好，還是差不多？ In your opinion, will your personal economic situation in the next 1 year become better, become worse, or remain about the same? (01) 會變非常不好 Will become very bad; (02) 會變不好 Will become bad; (03) 差不多 About the same; (04) 會變好 Will become good; (05) 會變非常好 Will become very good (06) 其他 Other; (96) 跳答 Skip question; (99) 遺漏值 Missing value.

# 2 Chinese Version of the Experimental Procedure

**Table A1:** Experimental Design and Procedure

| | Control Group | Treatment Group |
|---|---|---|
| **Agentic Participants** | 255 | 255 |
| **Pretreatment Question I** | **你听过美国前军事战略家杰克基恩说乌克兰战争是一项投资吗？ 美国只花 *660* 亿美元就能让乌克兰和俄罗斯打仗……基恩接着提到台湾也是一样，台湾公民是美国人打廉价战争的「投资」。美国冷酷精明，没有任何实际帮助台湾的意图！**<br>**Q1.** 请阅读以上陈述并评估其可靠性。(1) 非常不可靠 (2) 不可靠 (3) 未定 (4) 可靠 (5) 非常可靠 (96) 跳答<br>**Q2.** 阅读此陳述後，您有多大機會在社群媒體上分享？ (1) 極不可能 (2) 不可能 (3) 未定 (4) 可能 (5) 非常可能 (96) 跳答 | |
| **Pretreatment Question II** | **每年有打流感疫苗的要小心了！ 卫福部现在又大开后门，让高端引进国外低价疫苗，连研发都不用？这个无良政府为了赚钱，小老百姓可得睁大眼睛啊！ 出了事他们可不负责的事！**<br>**Q3.** 请阅读以上陈述并评估其可靠性。(1) 非常不可靠 (2) 不可靠 (3) 未定 (4) 可靠 (5) 非常可靠 (96) 跳答<br>**Q4.** 阅读此陳述後，您有多大機會在社群媒體上分享？ (1) 極不可能 (2) 不可能 (3) 未定 (4) 可能 (5) 非常可能 (96) 跳答 | |
| **Treatment: Debunk Message Exposure** | 無辟谣訊息 | **该陈述已通过独立事实查核人员的核查，并被证实为假新闻。** |
| **Post-treatment Question I** | **你听过美国前军事战略家杰克基恩说乌克兰战争是一项投资吗？ 美国只花 *660* 亿美元就能让乌克兰和俄罗斯打仗……基恩接着提到台湾也是一样，台湾公民是美国人打廉价战争的「投资」。美国冷酷精明，没有任何实际帮助台湾的意图！**<br>**Q5.** 請再次評估上述陳述的可靠性。(1) 非常不可靠 (2) 不可靠 (3) 未定 (4) 可靠 (5) 非常可靠 (96) 跳答<br>**Q6.** 阅讀此陳述後，您有多大機會在社群媒體上分享？ (1) 極不可能 (2) 不可能 (3) 未定 (4) 可能 (5) 非常可能 (96) 跳答 | **你听过美国前军事战略家杰克基恩说乌克兰战争是一项投资吗？ 美国只花 *660* 亿美元就能让乌克兰和俄罗斯打仗……基恩接着提到台湾也是一样，台湾公民是美国人打廉价战争的「投资」。美国冷酷精明，没有任何实际帮助台湾的意图！**<br>**Q5.** 在看过事实查核结果后，请再次评估上述陈述的可靠性。(1) 非常不可靠 (2) 不可靠 (3) 未定 (4) 可靠 (5) 非常可靠 (96) 跳答<br>**Q6.** 在看过事实查核结果后，阅读此陈述后，您有多大机会在社群媒体上分享？ (1) 極不可能 (2) 不可能 (3) 未定 (4) 可能 (5) 非常可能 (96) 跳答 |
| **Post-treatment Question II** | **每年有打流感疫苗的要小心了！ 卫福部现在又大开后门，让高端引进国外低价疫苗，连研发都不用？这个无良政府为了赚钱，小老百姓可得睁大眼睛啊！ 出了事他们可不负责的事！**<br>**Q7.** 請再次評估上述陳述的可靠性。(1) 非常不可靠 (2) 不可靠 (3) 未定 (4) 可靠 (5) 非常可靠 (96) 跳答<br>**Q8.** 阅讀此陳述後，您有多大機會在社群媒體上分享？ (1) 極不可能 (2) 不可能 (3) 未定 (4) 可能 (5) 非常可能 (96) 跳答 | **每年有打流感疫苗的要小心了！ 卫福部现在又大开后门，让高端引进国外低价疫苗，连研发都不用？这个无良政府为了赚钱，小老百姓可得睁大眼睛啊！ 出了事他们可不负责的事！**<br>**Q7.** 在看过事实查核结果后，请再次评估上述陈述的可靠性。(1) 非常不可靠 (2) 不可靠 (3) 未定 (4) 可靠 (5) 非常可靠 (96) 跳答<br>**Q8.** 在看过事实查核结果后，阅读此陈述后，您有多大机会在社群媒体上分享？ (1) 極不可能 (2) 不可能 (3) 未定 (4) 可能 (5) 非常可能 (96) 跳答 |

# 3 Chinese Version of the Exemplified Persona

```
{ "respondent_4436": {
        "name": "Voter ID_4436",
        "description": "Taipei City Resident",
        "system_message": """

    【基本資料】
    我是民國[65]年出生 (v39) ，女性 (v49)。最高學歷是[技術學院/大學 (06)]
    (v43)。我的總家庭收入在[90,000元至109,999元 (06)]之間
    (v48)。目前🄴籍在[台北市] (v40)。我父親是[本省閩南人 (02)] (v42)。

    【政治立場】
    我支持[民進黨 (01)] (v46)，對這個政黨的支持强度是[普通 (02)]
    (v47)。整體來说，我對政治的事情[有點感興趣 (03)]
    (v34)。對蔡英文政府的整體表現[滿意 (03)]
    (v35)。如果明天就要投票選總統，我會投給[賴清德]
    (v8)。如果明天就要投票選舉總統和立法委員，我[一定會 (04)]去投票
    (v13)。我平常[有點注意 (03)]媒體對美中臺關系的新聞報導 (v29)。

    【國家認同】
    我對身为臺灣人感到[非常光榮 (04)]
    (v36)。關於台灣的未來，我比較贊成[先維持現狀，以後再走向獨立 (02)] (v38)。
    當被問到認为自己是「台灣人」、「中國人」，
    或者都是時，我認为自己是[臺灣人 (01)] (v41)。

    【兩岸與國際关系觀點】
    我認为在未來10到20年，中國[可能 (03)]以武力攻打臺灣
    (v4s4)；而在未來5到10年，中國[不可能 (02)]以武力攻打臺灣
    (v4s5)。如果中國以武力攻打臺灣，我認为美國[可能 (03)]直接派兵援助臺灣
    (v7)。如果臺灣與中國發生戰争，我[會 (04)]抵抗
    (v32)。我對美國的整體印象[好 (03)] (v30)，對中國的整體印象[不好 (02)]
    (v31)。如果臺灣與中國發生戰争，我認为大多數臺灣人[會 (04)]抵抗 (v33)。

    【經濟評價】
    我認为目前美國與中國的軍事力量，[美國强一些
    (02)](v16)；而20年後，[中國强一些 (04)] (v17)。
    在經濟實力方面，目前[一樣强 (03)] (v18)；20年後，[中國强一些 (04)] (v19)。
    關於美國對台海的影響，在0到10的量表上，我給[7]分
    (v20)，較傾向認为美國促成台海兩岸穩定。
```

```
    關於中國對台灣的影響，在0到10的量表上，我給[3]分
    (v21)，較傾向認為中國對台灣的國家安全與民主自由威脅很大。
    關於台灣應該靠向哪一方，在0到10的量表上，我給[8]分
    (v22)，較傾向認為台灣應該靠向美國。

    Economic Performance Evaluation:
    我認為台灣現在的經濟狀況比1年以前[比較好 (04)]
    (v25)。我覺得自己目前的經濟狀況比1年以前[差不多 (03)]
    (v27)。對於台灣未來1年的經濟狀況，我認為[會變好 (04)]
    (v26)。而對於我個人未來1年的經濟狀🄴，我認為[差不多 (03)] (v28)。
"""}
}
```

# 4 Frequency Distribution of Key Variables in our Agent Respondents

**Figure A1:** Cross-Strait Unification

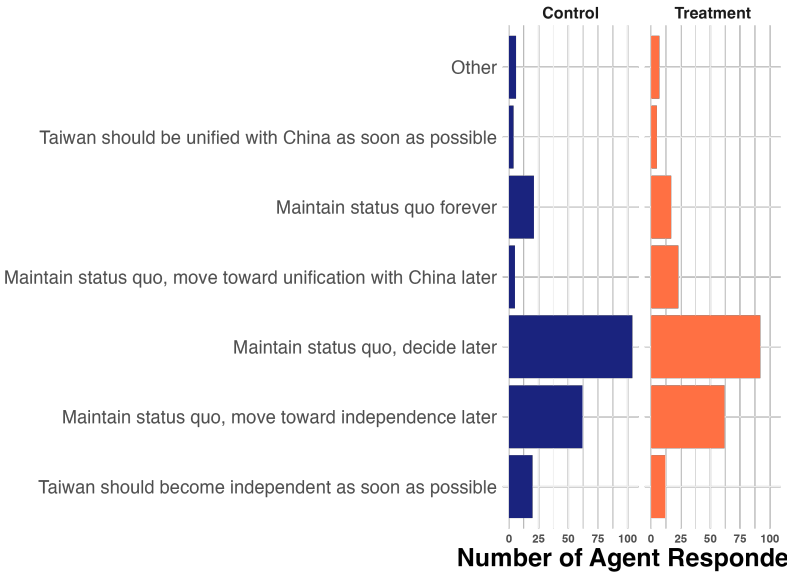



**Figure A2:** National Identity

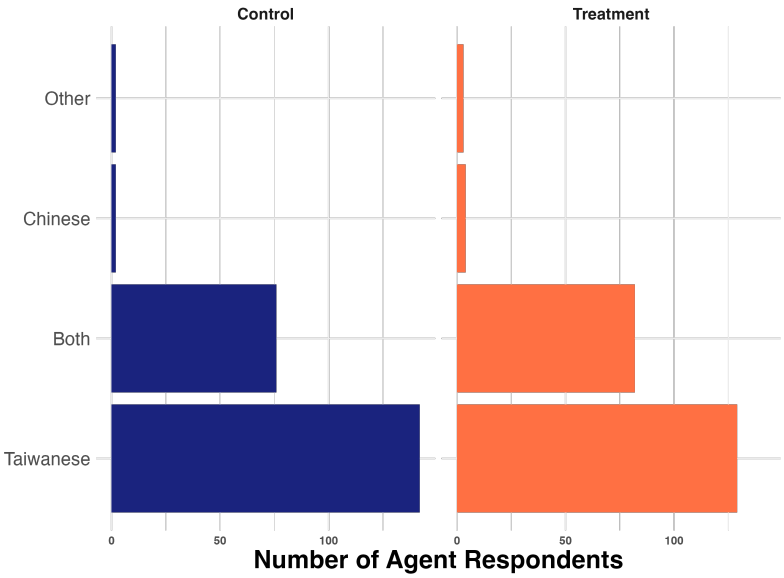


**Figure A3:** General Impression about China

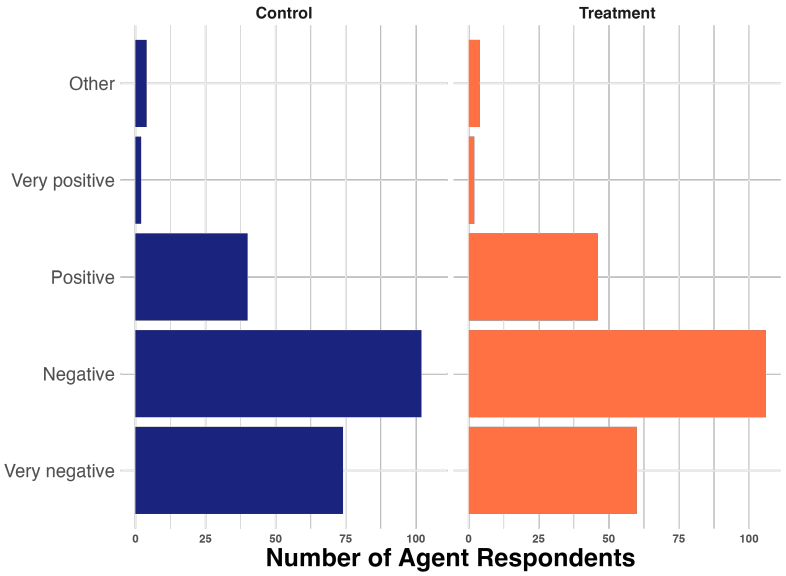



**Figure A4:** Resistance in a Potential War with China

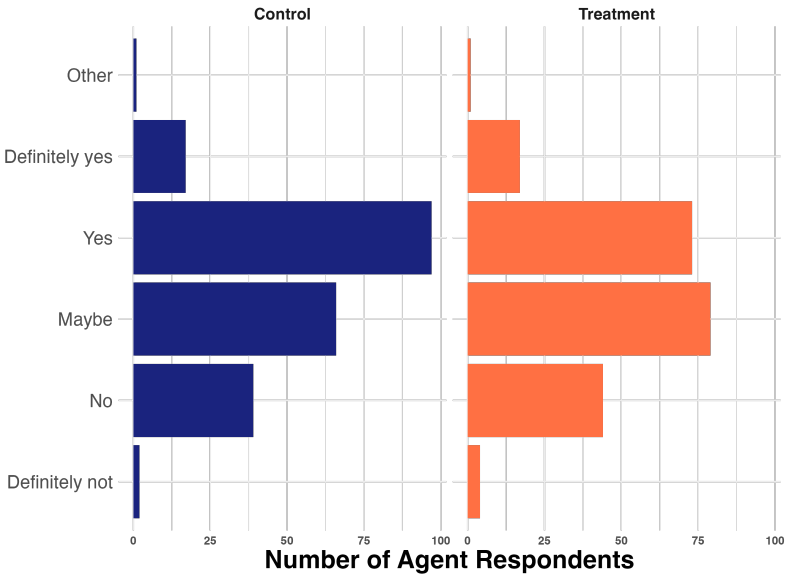


**Figure A5:** Party Affiliation