SURVEYPILOT: Finite-State Orchestrated Agentic Framework for Automated Human Opinion Collection from Social Media

Anonymous ACL submission

Abstract

Opinion survey research is a crucial method used by social scientists for understanding societal beliefs and behaviors. Traditional methodologies often entail high costs and limited scalability, while current automated methods such as opinion synthesis exhibit severe biases and lack traceability. In this paper, we introduce SURVEYPILOT, a novel finite-state orchestrated agentic framework that automates the collection and analysis of human opinions from social media platforms. SURVEYPILOT addresses the limitations of pioneering approaches by (i) providing transparency and traceability in each state of opinion collection and (ii) incorporating several techniques for mitigating biases, notably with a novel genetic algorithm for improving result diversity. Our extensive experiments reveal that SURVEYPILOT achieves a close alignment with authentic survey results across multiple domains, observing average relative improvements of 68.98% and 51.37% when comparing to opinion synthesis and agent-based approaches.

1 Introduction

Opinion survey research is a key method social scientists use to collect information about opinions, beliefs, and behaviors of a target group through formal interviews and questionnaires (Bryman, 2016; Bryson et al., 2012). Traditional surveys depend on labor-intensive approaches, such as phone interviews and web surveys, to collect data from statistically representative populations. A particularly concerning issue is the high cost and reliability of responses from human participants (Sun et al., 2024). In response to these challenges, this work explores *scalable*, *robust* and *verifiable* approaches to automatically collect responses and perform data analysis for pre-designed survey questions.

The recent advances utilize Large Language Models (LLMs) to generate synthetic responses that mirror human respondents by conditioning on demographic information, referring to as opinion synthesis (Ferraro et al., 2024) and social simulation (Chuang et al., 2024; Kamruzzaman and Kim, 2024). However, using LLMs as proxies for survey respondents presents three key limitations. First, LLMs exhibit inherent biases that skew responses toward certain demographic groups and harmless ones for sensitive survey topics (Sun et al., 2024). Second, LLMs are susceptible to hallucinations and prediction errors, and their opaque generation processes hinder traceability and error analysis, thereby raising accountability issues in social science research (Zhou et al., 2024; Gao et al., 2024). Third, LLMs have a fixed knowledge cutoff (typically 2023-2024), infeasible to capture recent demographic shifts or account for events occurring after their last training update (Sanders et al., 2023).

To address those limitations, we propose SURVEYPILOT, an LLM-based agentic framework that automatically collects human opinions from social media platforms. Our framework employs a workflow characterized by a finite-state machine (Carroll and Long, 1989) for orchestration and leverages a novel genetic algorithm designed to enhance opinion diversity - which is a key factor in reducing biases (Mehrabi et al., 2021). The agent operates in three key stages: (i) generating a diverse set of search queries derived from the given survey questions, (ii) identifying, filtering, and reranking relevant web pages from specified online sources, and (iii) extracting human opinions from these pages and representing them in a structured format for easy aggregation in survey responses. The genetic algorithm carefully balances query relevance and diversity to mitigate data bias, while the reranking component addresses the indexical bias present in search engine results. By sourcing *timely* real opinions, SURVEYPILOT naturally overcomes the knowledge cutoff constraint in LLMs and enhances interpretability, since each opinion is directly traceable to its original source and context.

We evaluate SURVEYPILOT through both extrinsic and intrinsic evaluations. For extrinsic evaluation, we replicate findings from established surveys (e.g. surveys of PEW Research Center) using SURVEYPILOT and compare its results with other automated approaches, measuring their correlation with actual human responses. For intrinsic evaluation, we assess the reliability of each state in SURVEYPILOT's workflow, demonstrating the effectiveness of our bias mitigation techniques. In summary, our key contributions are as follows:

- We present SURVEYPILOT, an LLM-driven framework for collecting authentic human opinions from social media, addressing the limitations of synthetic data generation.
- We demonstrate through extrinsic evaluation that SURVEYPILOT achieves higher correlations with human survey responses across multiple topics compared to existing social simulation and agent-based approaches. Specifically, when using LLAMA3.3-INSTRUCT 70B as the backbone model, **we observe average relative improvements of 68.98% and 51.37%** when comparing SURVEYPILOT to opinion synthesis and agent-based methods.
- We develop a genetic algorithm for search query diversification, with intrinsic evaluation validating its effectiveness in reducing bias and improving SURVEYPILOT's alignment with actual survey results.

2 SURVEYPILOT

Building on recent work in LLM-based agentic workflows for task-solving (Wu et al., 2024b; Shi et al., 2024), we introduce SURVEYPILOT, an agent-based system that systematically collects and analyzes human opinions from social media platforms. Given a survey question s from a questionnaire with a set of possible answer options $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$, SURVEYPILOT outputs a probability distribution:

$$P(\mathcal{O} \mid s) = \{P(o_i \mid s) \mid o_i \in \mathcal{O}\}, \qquad (1)$$

where each $P(o_i \mid s)$ represents the proportion of opinions supporting option o_i based on the collected data. Figure 1 shows the system operates as a finite-state machine with six key states:

Formatting Survey Question: The system extracts answer options from the survey question and paraphrases the options to guide later stages.

Search Query Generation: Through a genetic algorithm, the system generates diverse search queries to find relevant online discussions spanning multiple perspectives on each survey question.

Web Page Filtering: The system gathers web pages from Reddit and X/Twitter via the Google API, employing a reranking and filtering mechanism to mitigate indexical bias.

Opinion Gathering: Through automated browser interactions, the system collects opinions and their associated metadata from the filtered web pages.

Attribute Extraction: The system analyzes each opinion to extract key attributes such as language, gender, and their corresponding survey answer options, augmented by web searches when needed.

Evaluate Diversity: The system evaluates the distribution of opinions across different dimensions, ensuring comprehensive coverage of perspectives and providing feedback for refinement if necessary.

2.1 States of SURVEYPILOT

2.1.1 Formatting Survey Question

Our system focuses on categorical survey questions such as multiple-choice questions, yes-no questions, and rating scales to enable direct comparison with traditional survey results. For each question, the LLM agent extracts its response options and generates paraphrased alternatives (prompt templates in Appendix A.1). These variations prove crucial for the genetic algorithm in the query generation state, helping create diverse search queries and improve data coverage.

2.1.2 Search Query Generation

The search query generation stage takes the formatted question, its options, and paraphrases of options as input. This state is critical for ensuring the collected data represents diverse viewpoints across all possible survey responses. By diversifying search queries, the collected opinions will later then be diverse, which is a key factor in reducing biases (Mehrabi et al., 2021; Li et al., 2024). We prompt the LLM agent to generate search queries following specific criteria: maximizing perspective diversity, covering all survey options, and avoiding stance assumptions (prompt templates in Appendix B).

We frame this task as a query expansion problem in information retrieval (Piramuthu et al., 2013). Traditionally, an initial query $\mathcal{Q}_0$ retrieves a set of results $\mathcal{R}(\mathcal{Q}_0)$, which are then used to refine the query into an improved version $\mathcal{Q}^*$. Recent research has demonstrated that LLM-generated

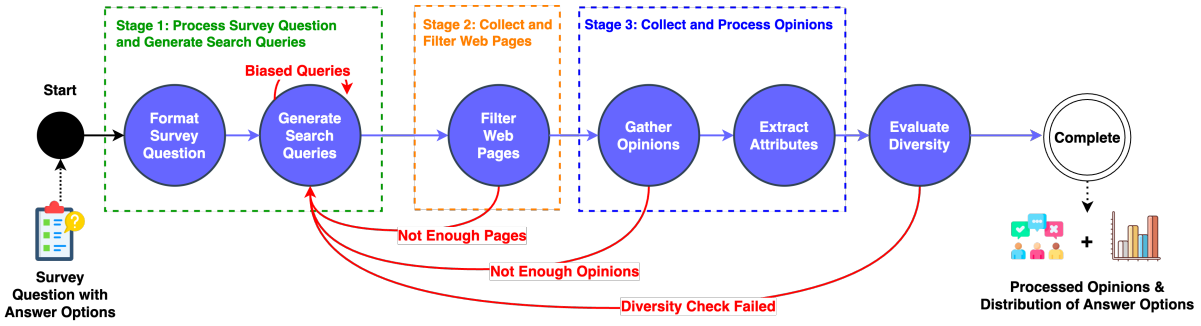

Figure 1: Illustration of SURVEYPILOT for opinion collection. SURVEYPILOT follows the finite-state machine design, takes a survey question as input and outputs processed opinions from social media. Red arrows indicates failure signals of the process.

search queries can effectively serve as query expansion strategies, improving coverage and precision (Wang et al., 2023a; Jagerman et al., 2023; Zhang et al., 2024). To diversify search queries, we integrate a genetic algorithm (GA) that iteratively optimizes the query set. Algorithm 1 outlines the process, which consists of four main stages: (i) *Population Initialization*, (ii) *Fitness Evaluation*, (iii) *Parent Selection*, and (iv) *Crossover and Mutation*.

Population Initialization. Let $\mathcal{Q}$ be the set of all possible search queries. For a survey question s, we use the LLM to generate n sets of candidate queries:

$$\mathcal{P}_0 = \{S_i \mid i = 1, 2, \ldots, n\}, \quad S_i \subset \mathcal{Q} \tag{2}$$

Each candidate set S_i undergoes iterative refinement through a Reflexion process (Shinn et al., 2023), producing a sequence:

$$S_i^{(0)} \rightarrow S_i^{(1)} \rightarrow \cdots \rightarrow S_i^{(N_{\text{Refine}})} \tag{3}$$

The final refined set $S_i^{(N_{\text{Refine}})}$ becomes part of the initial population $\mathcal{P}_0$.

Fitness Evaluation. For each candidate solution $S = \{q_1, q_2, \ldots, q_{|S|}\}$ in $\mathcal{P}_0$, we use a search engine to retrieve a set of web pages $R(q) \subset \mathcal{W}$ for each query $q \in S$, where $\mathcal{W}$ represents all available web pages. The diversity of S is measured by using the average overlapping rate between the search results (list of URLs) from each pair of queries:

$$\phi(S) = \frac{2}{|S|(|S|-1)} \sum_{\substack{q_i, q_j \in S \\ i<j}} \frac{|R(q_i) \cap R(q_j)|}{|R(q_i) \cup R(q_j)|}. \tag{4}$$

Lower values of $\phi(S)$ correspond to higher diversity. We define the fitness function as $\psi(S) = 1 - \phi(S)$, so that higher values of $\psi(S)$ indicate more desirable solutions.

Parent Selection. To evolve the population, we sample a fixed set of N_{Parents} candidate solutions from $\mathcal{P}_t$ for reproduction using Boltzmann Tournament Selection (Goldberg, 1990). Given the fitness $\psi(S)$ of each candidate $S \in \mathcal{P}_t$ at generation t, the probability of selecting S is given by

$$P(S) = \frac{\exp\left(\psi(S)/T\right)}{\sum_{S' \in \mathcal{P}_t} \exp\left(\psi(S')/T\right)}, \tag{5}$$

where $T > 0$ is a temperature parameter that regulates the selection pressure. This mechanism ensures that candidates with higher fitness are more likely to be chosen. We select exactly n parents for the next reproduction phase, which will undergo crossover and mutation.

Crossover and Mutation. Let $S^{(1)}$ and $S^{(2)}$ be two parent solutions selected from $\mathcal{P}_t$. The crossover operator combines these parents to generate an offspring solution S', which is denoted as $S' = \text{Crossover}\left(S^{(1)}, S^{(2)}\right)$. In our implementation, the LLM agent is directly prompted (see Appendix B for the prompt templates) to generate S' by merging features from both parents. Subsequently, a mutation operator is applied, wherein the LLM agent is further prompted to modify S' by substituting certain search queries with semantically related alternatives. The mutated offspring $\tilde{S}$ is then defined as $\tilde{S} = \text{Mutate}\left(S'\right)$, and $\tilde{S}$ is incorporated into the evolving population.

Final Selection. After a fixed number N_{Gen} of generations, we select the candidate solution S^* with the highest fitness:

$$S^* = \arg\max_{S \in \mathcal{P}_{N_{\text{Gen}}}} \psi(S). \tag{6}$$

In cases where multiple solutions achieve the maximal fitness, we construct a composite solution by

taking the union of their search query sets and removing duplicate queries. The hyperparameters of the GA (e.g. population size n, number of generations N_{Gen}) are specified in Appendix B.

Algorithm 1 Genetic Algorithm for Query Diversification

Input: Survey question s, LLM agent $\mathcal{A}$, population size n, parents number N_{Parents}, number of generations N_{Gen}, temperature T
Output: Diversified query set S^*

// Phase 1: Initialize Population
$\mathcal{P} \leftarrow \{ S_i = \text{Refine}(\mathcal{A}(s)) \mid i \in [n] \}$
// Phase 2: Evolution
for $g \leftarrow 1$ **to** N_{Gen} **do**
  // 2.1: Fitness Evaluation
  foreach $S \in \mathcal{P}$ **do**
    foreach $q \in S$ **do**
      $R_q \leftarrow \text{Search}(q)$
    $f(S) \leftarrow 1 - \text{Avg}\{\text{overlap}(R_i, R_j) \mid q_i, q_j \in S,\ i < j\}$
  // 2.2: Parent Selection
  $\mathcal{P}_{\text{sel}} \leftarrow \text{Select}_{N_{\text{Parents}}}(\mathcal{P}, f)$ $\mathcal{P}' \leftarrow \varnothing$
  // 2.3: Reproduction (Crossover & Mutation)
  foreach $(S_i, S_j) \in \mathcal{P}_{sel}$ **do**
    $S_{\text{child}} \leftarrow \text{LLM\_crossover}(S_i, S_j)$
    $S_{\text{child}} \leftarrow \text{LLM\_mutation}(S_{\text{child}})$
    $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{S_{\text{child}}\}$
  $\mathcal{P} \leftarrow \mathcal{P}'$
// Phase 3: Final Selection
$S_{\text{best}} \leftarrow \arg\max_{S \in \mathcal{P}} f(S)$ **if** *multiple* S_{best} *exist* **then**
  $S^* \leftarrow \text{Unique}\left(\bigcup S_{\text{best}}\right)$
else
  $S^* \leftarrow S_{\text{best}}$
return S^*

2.1.3 Web Page Filtering

With the generated search queries, we proceed to collect and filter web pages. At this stage, SURVEYPILOT focuses on retrieving posts and discussions from **Reddit and X/Twitter**, as these platforms host large, active communities with diverse user bases, providing a broad spectrum of opinions relevant to the survey question. Web page URLs are obtained using the Google API (Google, 2025). However, search engines often suffer from indexical bias (Mowshowitz and Kawaguchi, 2002), which can lead to suboptimal ranking of results.

Re-ranking Search Results. To mitigate indexical bias, we apply a re-ranking step before filtering. Specifically, we use the `BAAI/bge-reranker-v2-m3` model (Chen et al., 2024), a widely used multilingual pre-trained model for textual embedding in information retrieval. For each search query, we compute the cosine similarity between the query and the summarized content retrieved via the Google API. The web pages are then re-ranked from highest to lowest based on their relevance.

Filtering Search Results. After re-ranking, the top 200 results for each search query are selected based on their ranking scores. The page content is then converted to markdown format, and an LLM is prompted to assess its relevance to the survey question (see Appendix A.1 for prompt details). Finally, the filtered web pages from all queries are aggregated, and duplicate entries are removed to ensure a clean and diverse dataset.

2.1.4 Opinion Gathering

Using Playwright (Microsoft, 2025) to simulate real user interactions, our system use LLMs to comprehensively collects opinions related to the survey question from Reddit posts and X/Twitter threads. For each URL, the system extracts detailed information including the opinion-related text such as comments on Reddit posts and tweets on X, along with details such as the publication timestamp, username, and any associated metadata provided in the HTML or embedded data structures. This method preserves the original context of each opinion, which is essential for later stages where we extract further attributes and evaluate the diversity of the collected data. Each opinion is stored in JSON format for subsequent analysis.

2.1.5 Attribute Extraction

In this state, we extract attributes for comparing opinion distributions with those from actual surveys. For each collected opinion, we extract: (i) gender, (ii) language, and (iii) the list of preferred answer options expressed in the opinion. Recall that in the first state - Formatting Survey Question (Section 2.1.1), we have extracted the answer options from the survey question. In this state, the LLM agent is instructed to determine whether each opinion's content mentions or implies any answer options (prompt templates shown in the Appendix A.1). For instance, if the question is "What is your favourite movie genre?", the opinion text could mention several genres of preferences. Additionally, the agent collects information on languages and genders (if expressed in the opinion).

After the required attributes are extracted, we can compute the output probability distribution $P(\mathcal{O} \mid s)$ of all options $\mathcal{O}$ (denoted in Section 2)

by calculating the frequency of each option in the collected data.

2.1.6 Diversity Ensurance

After the human opinions are collected and processed, a diversity check is performed by executing several metrics. In this state, we propose several conditions for diversity checking:

Language Coverage. If the survey question is intended for participants of multiple countries or regions, we compute the frequencies of languages represented in the data to see if the data is dominant by any language (i.e. having its frequency higher than 50%). If there is a dominant language, this condition is failed.

Gender Coverage. For survey questions intended for both genders, we make sure that the data includes at least 100 opinions for each gender to pass this condition.

Answer Option Coverage. It is also important for the data to cover all answer options (eg. viewpoints, preferences) required by the survey questions. The data needs at least 50 opinions for each answer option to pass this condition.

Diversity checking is considered passed if all of the above conditions are passed. Otherwise, corresponding feedback to the failed conditions is provided to the agent so that the solution can be improved in the next iteration (starting at the Search Query Generation state).

2.2 Agent Configuration

2.2.1 Memory Management

We follow RecAgent (Wang et al., 2024) to employ a dual-tiered memory management mechanism for our agent, where it differentiates between temporary memory and main memory (long-term memory).

Temporary Memory. The temporary memory is allocated for high-speed, repetitive tasks such as filtering web pages (Section 2.1.3) and processing human opinions (Section 2.1.5). For each sample (either the content of a web page or human opinion), temporary memory records the response of the agent and the corresponding outcome. For instance, in the case of extracting attributes from an opinion, the response of the agent is a JSON-formatted response, where each key represents a required attribute in the instruction. Correspondingly, the outcome is either a *"Successful"* flag or an error message indicating an error in parsing the response of the agent, so that the agent knows to correct its response. Temporary memory is used only in three states - web page filtering, opinion gathering and attribute extraction, and it is re-initialize when the agent enters one of the three states above.

Main Memory. In contrast, the main memory serves as a persistent repository where crucial actions from each state and feedback are recorded, enabling the agent to maintain long-term context and learn from past interactions. As illustrated in Figure 1, each error message (indicated as a red arrow) is represented as feedback and saved to the main memory. Outcomes of every main action taken by the agent in each state are also recorded and provided as context for the agent to act in the next state.

Memory Representation. We represent each type of memory as a list of actions, outcomes, and feedback consecutively. Before each action of the agent, it is provided with either the temporary memory or main memory in the Markdown format, with headers indicating the roles of interaction (i.e. *# Agent's Action:*, *# Outcome:*, *# User's Feedback:*).

2.2.2 Prompting Strategy

Except for the search query generation state (Section 2.1.2), which uses Reflexion (Shinn et al., 2023) for better creativity in the responses, for other states, we apply the ReAct prompt format (Yao et al., 2023) for the LLM agent due to its faster execution compared to Reflexion. Specifically, the agent is instructed to provide its thoughts on a problem and then decide on its action before responding and observing the results of the action taken. The contents from either the main memory or temporary memory are included as the context in the prompt.

3 Experiments

In our experiments, we investigate to what extent SURVEYPILOT and other opinion synthesis approaches can replicate results from established surveys. Additionally, we evaluate the effectiveness of each states of SURVEYPILOT to justify our design choices.

3.1 Model Configuration

For all LLMs, we set the temperature $t = 0.5$ and $top\_p = 0.95$ for inference. LLMs tested in our experiments include QWEN2.5-INSTRUCT

14B and 32B (Qwen et al., 2025), LLAMA3.1-INSTRUCT 8B, and LLAMA3.3-INSTRUCT 70B, where LLAMA3.3-INSTRUCT 70B is used by default for SURVEYPILOT. Appendix F shows the deployment setups of models.

3.2 Baseline Configuration

Opinion Synthesis Baseline. Following other works on opinion synthesis (Ferraro et al., 2024; Chuang et al., 2024; Kamruzzaman and Kim, 2024; AlKhamissi et al., 2024), we employ a prompt template to assign personalities to LLMs and conduct surveys by prompting LLMs to answer survey questions. The prompt template is given in Appendix A.2. For each survey question, we sample 10,000 responses from each LLM.

Agentic Frameworks for Data Collection. We choose to evaluate ScrapeGraphAI (Marco Perini, 2024) and AutoGen (Wu et al., 2024a). These frameworks are instructed to collect and process human opinions from Reddit and X, and the collected data will be compared with actual surveys. Details of setups for ScrapeGraphAI and AutoGen are given in Appendix C.

3.3 Datasets

We choose 40 survey questions from surveys that are conducted in late 2023 and 2024 for the experiments. Widely-used surveys such as the World Values Survey (Ing, 2022) or the GLOBE Survey (House et al., 2004) are not included, as these may have appeared in the pre-training data of LLMs (Appendix D.1 shows the data contamination experiments of survey datasets). Our chosen survey questions cover 4 topics: (i) *technology*, (ii) *entertainment*, (iii) *cuisine*, and (iv) *religion*. Most of the questions are taken from surveys of the PEW Research Center[1]. Details of the survey questions and their sources are given in Appendix E.

3.4 Evaluation Metric

To compare data collected or generated by different methods with results from actual surveys, we follow (Durmus et al., 2024; Sorensen et al.) to use Jensen-Shannon divergence as our evaluation metric. Jensen-Shannon divergence takes two vectors as input, which are the answer distribution from surveys and the data distribution from methods. Lower divergence values indicate a better correlation between two distributions.

[1]https://www.pewresearch.org/

3.5 Extrinsic Evaluation

3.5.1 Main Results

The performance of different methods in survey result recreation is shown in Table 1. Agent-based methods such as SURVEYPILOT, AutoGen, and ScrapeGraphAI outperform the opinion synthesis method on all domains. SURVEYPILOT consistently achieves lower divergence values over other methods in all domains, in which using LLAMA3.3-INSTRUCT 70B as the backbone model achieves the optimal results. Notably, the opinion synthesis approach with different LLMs performs worse in every domain, with divergence values more than triple those of SURVEYPILOT in several cases. In terms of other agentic frameworks, SURVEYPILOT achieves average relative improvements of 44% and 58% when comparing to ScrapeGraphAI and AutoGen, respectively. ScrapeGraphAI, as being designed for data scraping, also has better results than AutoGen. These results have demonstrated the effectiveness of SURVEYPILOT compared to other approaches.

When switching to different LLMs as backbone models, there are performance drops of SURVEYPILOT as the model size is progressively smaller. LLAMA3.1-INSTRUCT 8B achieves the worst performance. This is mainly due to the search query generation step, where larger models like LLAMA3.3-INSTRUCT 70B and QWEN2.5-INSTRUCT 32B can generate more diverse and natural queries than smaller models. Appendix D.4 shows the differences between the search queries generated by different LLMs.

3.5.2 Analysis on Demographic-Specific Surveys

In this section, we focus on evaluating methods on demographic-specific survey questions to demonstrate the abilities of SURVEYPILOT in reducing biases. in for evaluation. Table 2 shows the results on country-specific and gender-specific survey questions. Opinion synthesis continues to have the worst results, notably with heavy bias towards both the US distributions and male distributions, where there are much lower divergence values in these columns. Regarding other agentic frameworks, they have fair representation between genders, but fails to address the problem of country-level bias - Italy and Spain still have substantially larger divergence values compared to other countries. By incorporating de-biasing techniques, SURVEYPI-

Method	Alignment with Domains (Jensen–Shannon divergence ↓)			
	Technology	Entertainment	Cuisine	Religion
Opinion Synthesis with Assigned Personalities				
- LLAMA3.1-INSTRUCT 8B	0.65	0.61	0.77	0.72
- QWEN2.5-INSTRUCT 14B	0.59	0.47	0.67	0.71
- QWEN2.5-INSTRUCT 32B	0.55	0.51	0.46	0.73
- LLAMA3.3-INSTRUCT 70B	0.49	0.31	0.38	0.65
SURVEYPILOT (Ours)				
- LLAMA3.1-INSTRUCT 8B	0.29	0.33	0.28	0.28
- QWEN2.5-INSTRUCT 14B	0.23	0.20	0.18	0.19
- QWEN2.5-INSTRUCT 32B	0.18	0.12	**0.15**	0.15
- LLAMA3.3-INSTRUCT 70B - **default setting**	**0.15**	**0.11**	**0.15**	**0.12**
Other Agentic Frameworks				
AutoGen (Wu et al., 2024a) - LLAMA3.3-INSTRUCT 70B	0.39	0.21	0.28	0.55
ScrapeGraphAI (Marco Perini, 2024) - LLAMA3.3-INSTRUCT 70B	0.30	0.21	0.19	0.29

Table 1: Performance of different methods in survey result recreation (Jensen–Shannon divergence ↓). **Blue** highlights the best results from each column.

Model	Country-Specific Results					Gender-Specific Results	
	America	France	Italy	Spain	Germany	Male	Female
Opinion Synthesis	0.33	0.46	0.58	0.51	0.42	0.52	0.59
SURVEYPILOT (Ours)	**0.12**	**0.14**	**0.14**	**0.12**	**0.13**	**0.18**	**0.18**
Other Agentic Frameworks							
AutoGen (Wu et al., 2024a)	0.24	0.29	0.31	0.35	0.26	0.36	0.38
ScrapeGraphAI (Marco Perini, 2024)	0.24	0.24	0.35	0.31	0.25	0.21	0.20

Table 2: Performance of different methods on country-specific and gender-specific survey questions (Jensen–Shannon divergence ↓). **Blue** highlights the best results from each column. LLAMA3.3-INSTRUCT 70B is used as the backbone LLM in this experiment.

LOT continues to achieve optimal performance in both country-specific and gender-specific surveys. Appendix D.2, table 7 shows that when removing the proposed GA and page reranker, our SURVEYPILOT achieves similar results to AutoGen and ScrapegraphAI, suggesting the effectiveness of debiasing techniques.

3.6 Intrinsic Evaluation

3.6.1 Independent Evaluation of Each State in SURVEYPILOT

Process	Accuracy
Formatting Survey Questions	98.00
Web Page Filtering	96.00
Opinion Gathering	96.67
Attribute Extraction	94.00

Table 3: Performance of SURVEYPILOT in several processes (Accuracy). The backbone LLM in this experiment is LLAMA3.3-INSTRUCT 70B.

We measure the performance of SURVEYPILOT when executing the following processes: (i) Formatting Survey Questions, (ii) Web Page Filtering, (iii) Opinion Gathering, and (iv) Attribute Extraction. For each process, we manually label a test set and measure the performance of the agent using accuracy score (details are given in Appendix D.3). Table 3 shows the performance of SURVEYPILOT in each of the above process. As these tasks are straightforward, the accuracy scores are consistently high in every process.

3.6.2 Design Choices Justification

We justify our design choices of SURVEYPILOT in the search query generation and web page filtering states by comparing the results of proposed components with other alternatives. Specifically, for the web page filtering state, we experiment with changing to: (i) use BertScore (Zhang et al., 2020) as page reranker and (ii) removing the reranking step. The BertScore receives the same inputs as our default reranker (Section 2.1.3), which are the search query and summarized page contents, and produces similarity scores. Regarding the query generation step, we experiment with substituting the GA with the following methods:

- **Best-of-N** (Brown et al., 2025). We sampled 1000 sets of search queries by prompting the LLM agent and apply the fitness function of the GA to take out the best solution.
- **Self-Consistency** (Wang et al., 2023b). Simi-

Method	Alignment with Domains (Jensen–Shannon divergence ↓)			
	Technology	Entertainment	Cuisine	Religion
SURVEYPILOT				
- Default Settings (with GA and BGE Page Reranker)	**0.15**	**0.11**	**0.15**	**0.12**
Other search query generation methods				
- Best-of-N (Brown et al., 2025)	0.18	0.23	0.18	0.21
- Self-Consistency (Wang et al., 2023b)	0.20	0.21	0.19	0.27
- 1-Pass	0.27	0.29	0.23	0.28
Other web page reranking methods				
- BertScore (Zhang et al., 2020) as page reranker	0.18	0.20	0.18	0.21
- Without page reranker	0.25	0.19	0.26	0.27

Table 4: Performance of SURVEYPILOT when substituting proposed techniques with alternative methods (Jensen–Shannon divergence ↓). **Blue** highlights the best results from each column.

lar to Best-of-N, 1000 sets of search queries are sampled. Then, queries with frequencies more than 100 are kept in the final solution.

- **1-Pass.** Here, one single set of search queries is generated by the LLM agent.

Experimental results are shown in Table 4. Using our proposed GA for query generation and page reranker model consistently yields optimal results across all domains. This directly supports our design choices, as each alternative approach increases divergence, indicating weaker alignment to survey distributions. Among the query generation strategies, the simplest 1-Pass performs the worst. Best-of-N and Self-Consistency show moderate performance, indicating that selecting one solution without refinements leads to higher divergence than our GA-based approach. By contrast, the GA iteratively refines a population of search queries with fitness-driven updates, leading to richer coverage and more comprehensive retrieval of viewpoints.

Regarding the page reranker, replacing the BGE Reranker with BertScore results in higher divergence values, indicating that BGE's page-level embeddings capture semantic similarity more effectively. Additionally, omitting the reranking step substantially degrades performance due to search engine indexical bias. Additional experiments on country- and gender-specific survey questions are presented in Appendix D.2.

4 Related Works

Opinion Survey Research with LLMs. LLMs have been shown to mimic human survey responses by adopting personas and generating opinions that align with traditional results (Ferraro et al., 2024; Chuang et al., 2024; Kamruzzaman and Kim, 2024). (Yeykelis et al., 2024) demonstrate that opinion synthesis with persona-based prompts can replicate media effects studies, while (Kim and Lee, 2024) fine-tunes an LLM to predict missing responses. Despite reducing cost and time, these approaches face challenges from knowledge cutoff constraints (Sanders et al., 2023), training data biases (Kamruzzaman and Kim, 2024; Naous et al., 2024) and lack traceability. SURVEYPILOT overcomes these issues by collecting authentic and diverse social media opinions.

Agentic Frameworks for Data Collection. Recent works have automated data collection with LLM-based agents. ScrapeGraphAI (Marco Perini, 2024) uses LLMs and graph logic to interact with browsers, while frameworks like AutoGen (Wu et al., 2024a), MetaGPT (Hong et al., 2024), and SmolAgents (Roucher et al., 2025) serve broader purposes. SURVEYPILOT differentiates itself from other agentic frameworks by incorporating several techniques such as search query diversification and web page reranking to reduce the bias in data distribution.

5 Conclusions

This paper presented SURVEYPILOT, an agentic framework that overcomes traditional and automated survey limitations. By integrating a genetic algorithm for search query diversification with robust web page filtering and opinion extraction, SURVEYPILOT captures diverse, authentic human opinions from social media. Experimental results show that SURVEYPILOT outperforms both opinion synthesis approaches and other agent-based frameworks, yielding substantial lower divergence from actual survey responses. Future works will broaden data sources and further reduce bias. In summary, SURVEYPILOT represents a significant step forward in opinion survey research, offering a *robust*, *scalable*, and *verifiable* solution.

Limitations

When collecting human opinions using SURVEYPILOT and other agentic frameworks, we cannot use demographic information of users (e.g. nationalities, age groups...) on social media platforms, even if that information is publicly available, as it may breach data privacy laws and the platforms' terms of service. Hence, we cannot perform deeper analyses or experiments that depend on demographic information.

Ethical Considerations

This study was reviewed and approved by the Ethics Review Board of our organization. To regulate the use of SURVEYPILOT, we outline several ethical considerations and emphasize potential risks.

Misuse of SURVEYPILOT. The primary goal of SURVEYPILOT is to help social scientists in collecting and analyzing human opinions on social media for specific survey questions. The collected data by SURVEYPILOT may contain some content that could be perceived as unsafe or harmful, particularly when receiving controversial survey question, such as discrimination towards specific groups. SURVEYPILOT is released solely for academic and research purposes. Any form of misuse, including employing this framework to collect harmful content, is strictly prohibited. Users are expected to adhere to the highest ethical standards, ensuring the responsible use of SURVEYPILOT in alignment with research ethics. The authors and creators of SURVEYPILOT hold no liability for misuse, misinterpretation, or unintended consequences of the framework.

Potential Bias. While SURVEYPILOT has several bias reduction techniques, it is impossible to eliminate bias entirely. For instance, social media opinions may have inherent biases, as certain groups may have not used social media to express their opinions, leading to skewed distribution of the collected opinions. Therefore, the framework should be viewed as a tool for social science research and improvement rather than final solutions to the task of automated opinion survey research. By releasing SURVEYPILOT, we aim to contribute to the responsible development of AI technologies for the field of social science. All users of this framework are expected to use it under ethical research practices, ensuring transparency and fairness.

References

2022. World values survey: All rounds – country-pooled datafile. Dataset.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Bradley Brown, Jordan Juravsky, Ryan Saul Ehrlich, Ronald Clark, Quoc V Le, Christopher Re, and Azalia Mirhoseini. 2025. Large language monkeys: Scaling inference compute with repeated sampling.

Alan Bryman. 2016. *Social research methods*. Oxford university press.

Gregory L Bryson, Alexis F Turgeon, and Peter T Choi. 2012. The science of opinion: survey methods in research. *Canadian Journal of Anesthesia*, 59(8):736.

John Carroll and Darrell Long. 1989. *Theory of finite automata with an introduction to formal languages*. Prentice-Hall, Inc., USA.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico. Association for Computational Linguistics.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. 2024. Agent-based modelling meets generative ai in social network simulations. *Preprint*, arXiv:2411.16031.

C. Gao, X. Lan, N. Li, et al. 2024. Large language models empowered agent-based modeling and simulation:

a survey and perspectives. *Humanit Soc Sci Commun*, 11:1259.

David E. Goldberg. 1990. A note on boltzmann tournament selection for genetic algorithms and population-oriented simulated annealing. *Complex Syst.*, 4.

Google. 2025. Google api documentation. Accessed: 2025-02-08.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Robert J. House, Peter J. Hanges, Mohammad Javidan, Paul W. Dorfman, and V. Gupta. 2004. *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*. Sage Publications, Thousand Oaks, CA, USA.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *Preprint*, arXiv:2305.03653.

Mahammed Kamruzzaman and Gene Louis Kim. 2024. Exploring changes in nation perception with nationality-assigned personas in llms. *Preprint*, arXiv:2406.13993.

Junsol Kim and Byungkyu Lee. 2024. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *Preprint*, arXiv:2305.09620.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe, and Chris Callison-Burch. 2024. Bordirlines: A dataset for evaluating cross-lingual retrieval-augmented generation. *Preprint*, arXiv:2410.01171.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Marco Vinciguerra Marco Perini, Lorenzo Padoan. 2024. Scrapegraph-ai. A Python library for scraping leveraging large language models.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Microsoft. 2025. Playwright.

Abbe Mowshowitz and Akira Kawaguchi. 2002. Assessing bias in search engines. *Information Processing Management*, 38(1):141–156.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. 2013. Visual search: A large-scale perspective. In *Handbook of Statistics*, pages 269–297. Elsevier.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. 'smolagents': a smol library to build great agentic systems. `https://github.com/huggingface/smolagents`.

Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the Potential of AI-based Political Issue Polling. *Harvard Data Science Review*, 5(4). Https://hdsr.mitpress.mit.edu/pub/dm2hrtx0.

Lu Shi, Bin Qi, Jiarui Luo, Yang Zhang, Zhanzhao Liang, Zhaowei Gao, Wenke Deng, and Lin Sun. 2024. Aegis:an advanced LLM-based multi-agent for intelligent functional safety engineering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1571–1583, Miami, Florida, US. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*.

Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. 2024. Random silicon sampling:

Simulating human sub-population opinion using a large language model based on group-level demographic information. *Preprint*, arXiv:2402.18144.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2024. User behavior simulation with large language model-based agents. *Preprint*, arXiv:2306.02552.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024a. Autogen: Enabling next-gen LLM applications via multi-agent conversation.

Yiran Wu, Tianwei Yue, Shaokun Zhang, Chi Wang, and Qingyun Wu. 2024b. Stateflow: Enhancing LLM task-solving through state-driven workflows. In *First Conference on Language Modeling*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Leo Yeykelis, Kaavya Pichai, James J. Cummings, and Byron Reeves. 2024. Using large language models to create ai personas for replication and prediction of media effects: An empirical test of 133 published experimental research findings. *Preprint*, arXiv:2408.16073.

Le Zhang, Yihong Wu, Qian Yang, and Jian-Yun Nie. 2024. Exploring the best practices of query expansion with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1872–1883, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714, Miami, Florida, USA. Association for Computational Linguistics.

A Prompt Templates

A.1 Prompt Templates of SURVEYPILOT

```
Given the following survey question:
{survey_question}

Find human opinions on Reddit or X/Twitter to
answer the question.
```

Listing 1: The user requirements used in all prompt templates of SURVEYPILOT.

In this section, we present the default prompt templates used in each state of SURVEYPILOT, including:

Formatting Survey Question. To format survey question, we use 2 prompt templates. First, we use the prompt in Listing 4 to extract the categories or options given in the survey question. Then, we apply the prompt in Listing 5 to instruct the LLM agent to generate multiple alternatives of each category by paraphrasing it.

Search Query Generation. The search query generation prompt is given in Listing 7. This prompt template takes the user requirements as input, and it is use in the Population Initialization phase of the genetic algorithm.

Web Page Filtering. To filter the relevant web pages after reranking, we prompt the LLM agent with the prompt template in Listing 6. This prompt template takes the user requirements, the title and content of a page as inputs.

Opinion Gathering. For each web page, the LLM agent is instructed to iteratively collect the opinions from the page in the first page load, and scrolls until the end of the page. The prompt template is shown in Listing 10, where we have the agent decide to return a "Scroll" or "Terminate" flag, and the Playwright library will handle the flag. The agent is also provided with their last collected opinions, so they can easily decide whether to terminate the process.

Attribute Extraction. To extract the answer options from the collected human opinions, we use the prompt template in Listing 9. The template takes the user requirements, the source of the post, the post content, and the opinion content as inputs. Regarding the extraction process of genders and languages, we use the prompt in Listing 8.

In all of our prompt templates, there is the `<user-requirements>` parameter as input. The `<user-requirements>` is described in Listing 1.

A.2 Prompt Template of Opinion Synthesis

We experiment with different prompt templates that assign LLMs with personas and keep the best performing one in our experiment. We use a similar prompt to that of (AlKhamissi et al., 2024) as the opinion synthesis prompt in our experiments, where we assign LLMs with a nationality and gender, and ask models to answer survey questions based on the given personas. The prompt is given in Listing 3.

B Genetic Algorithm Configuration

B.1 Prompt Templates

Recall in Section 2.1.2, we use the LLM agent in several processes of the genetic algorithm. Listing 15 shows the prompt template in the refinement process of Population Initialization. Specifically, the LLM agent acts as both the reviewer and executor with two different system prompts and iteratively refine its solution. Regarding the Crossover and Mutation processes, Listing 14 describes the prompt templates used.

B.2 Hyperparameters

We empirically set the hyperparameters of the GA to achieve a balance between diversity and computational cost. Values for each parameter are given in Table 5.

Param	Description	Value
n	Population size	50
N_{Parents}	Number of parents	10
N_{Gen}	Number of generations	3
T	Temperature	0.1
N_{Refine}	Number of refinement iterations	5

Table 5: Hyperparameter values of the genetic algorithm in SURVEYPILOT.

C Agentic Framework Configuration

C.1 ScrapeGraphAI Configuration

To configure the ScrapeGraphAI (Marco Perini, 2024) framework for our experiments, we use the OmniSearchGraph module for orchestrating the web page collection process, and use the OmniScraperGraph for browsing and collecting human opinions on the collected web pages.

Collecting Web Pages. To collect relevant web pages to the survey questions, we use the OmniSearchGraph module with the prompt in Listing 2. For this module, we set the maximum search

results to be 200, which is identical to that of SURVEYPILOT.

```
Given the following survey question:
{survey_question}

Find web pages that contain discussions that are
 relevant to answer the survey question.

Please remember to:
- Provide the page URLs in a single list like
this: [\"url 1\", \"url 2\",...]
- Only response with the list of URLs, do not
response with anything else.
- The web pages should be Reddit or X/Twitter
posts discussing the topics or problems that are
 mentioned in the survey question.
```

Listing 2: Web page collection prompt of ScrapeGraphAI.

Collecting & Processing Human Opinions. To collect human opinions from the collected web pages, we apply the OmniScraperGraph, which has PlayWright integrated and enables the framework to scroll a web page until the end. The prompt template for this task is shown in Listing 11, which is a combined template of opinion gathering and atribute extraction prompts of SURVEYPILOT. The web page HTML content is given to ScrapeGraphAI to extract all of the opinions in that page.

C.2 AutoGen Configuration

To configure the AutoGen (Wu et al., 2024a) framework for our experiments, we follow their principle design by using multiple agents as a group for collaboration, each having a specific role. The agents used in AutoGen are as follows:

User Proxy Agent. This agent is used to handle the user prompt, as well as to return the results of tool calls made by the Executor Agent.

Manager Agent. This agent is the most important one in AutoGen. It is used to orchestrate the workflow of the group according to the plan given by the user. Specifically, the Manager Agent choose the next agent to "speak" in the discussion of the group.

Planner Agent. Planner Agent decides whether a step in the plan is executed successfully before moving on to the next step. It is also used to inform other agents about the current step, helping other agents remember what the plan is by repeating it.

Executor Agent. This agent is informed by the Planner Agent about the current step and what tasks need to be done. It then execute the tasks and use the provided tools if applicable.

Reviewer Agent. Reviewer Agent helps to review the responses of the Executor Agent. This agent is in charge of providing feedbacks to the Executor Agent, whether it is to improve the response quality (eg. improving the diversity of generated search queries) or informing about errors in response format or tool calling.

The user prompt to AutoGen is given in Listing 12, while Listing 13 shows the system prompts of all agents. Since AutoGen does not have sufficient tools to collect human opinions from social media, we integrate it with (i) Google API for collecting relevant web pages and (ii) PlayWright functions to browse the web pages and collect opinions.

D Additional Experiments

D.1 Data Contamination Analysis

Model	WOS	GLOBE	Our Data
LLAMA3.3 70B	19.72	15.01	3.89
QWEN2.5 32B	16.96	13.24	3.27
QWEN2.5 14B	14.61	14.47	1.69
LLAMA3.1	17.24	12.63	1.20

Table 6: Average data contamination rate of different LLMs on survey datasets.

To show the data contamination rate of established surveys (eg. World Value Survey - WOS and GLOBE), we use a method similar to (Liu et al., 2024), which asks LLMs to complete a sample provided with a truncated version of that sample. Specifically, we truncate the questions in a survey dataset and feed them as input to LLMs for completion. Additionally, we use the base version of LLMs instead of the instruction fine-tuned version to have a more accurate representation of the data contamination problems in the pre-training data of these models.

Finally, we compute the length of the longest common sequence (LCS) between the prediction of LLMs and the ground truth completion. The sequence here is defined as a sequence of words. The data contamination rate of one test sample (the survey question) is then calculated as the ratio of the length of LCS and the length of the ground truth prediction.

Table 6 shows the average data contamination rate of different LLMs on the WOS, GLOBE, and

Model	Country-Specific Results					Gender-Specific Results	
	America	France	Italy	Spain	Germany	Male	Female
SURVEYPILOT							
- Default Settings (with GA and BGE Page Reranker)	**0.12**	**0.14**	**0.14**	**0.12**	**0.13**	**0.18**	**0.18**
Other search query generation methods							
- Best-of-N (Brown et al., 2025)	0.14	0.14	0.16	0.15	0.14	0.20	0.19
- Self-Consistency (Wang et al., 2023b)	0.17	0.18	0.22	0.17	0.19	0.20	0.23
- 1-Pass	0.20	0.26	0.28	0.25	0.25	0.25	0.30
Other web page reranking methods							
- BertScore (Zhang et al., 2020) as page reranker	0.17	0.19	0.18	0.22	0.16	0.23	0.26
- Without page reranker	0.19	0.19	0.23	0.21	0.21	0.20	0.26

Table 7: Performance of SURVEYPILOT when substituting proposed techniques with alternative methods on country-specific and gender-specific survey questions. **Blue** highlights the best results from each column. LLAMA3.3-INSTRUCT 70B is used as the backbone LLM in this experiment.

our collected survey questions. Our collected surveys have the lowest contamination rate on different LLMs, with a large gap compared to WOS and GLOBE. Notably, as the size of models grows, they can remember more of their training data, hence resulting in higher contamination rates on WOS and GLOBE. Due to the potential of contamination of WOS and GLOBE, we opt to not include these datasets in our list of survey questions.

D.2 Addtional Results for Design Justification

Table 7 shows additional experimental results to support our design choices. Among other methods for search query generation, Best-of-N achieves the closest divergence values with the proposed genetic algorithm. 1-Pass continues to perform the worst, with biases appear in the American distribution and Male distribution.

Regarding other web page reranking methods, using BertScore exhibits a substantial decrease in performance, and also shows bias to the US culture. Without any page reranker, the framework performs much worse in gender-specific questions. These results have justified our uses of the proposed genetic algorithm and page reranker in reducing the biases when collecting human opinions.

D.3 Additional Information in Intrinsic Evaluation

In this section, we describe the labeling and evaluation processes to evaluate the performance of the LLM agent in each state of SURVEYPILOT.

Formatting Survey Questions. We evaluate the LLM agent by calculating the accuracy of extracting categories (opinions) given in the 40 collected survey questions. One accurate extraction is considered as when all of the categories in a question are extracted, with the exact same categorical names.

Web Page Filtering. For each of the 40 survey questions, we sampled 5 web pages and manually labeled them for this evaluation, resulting in a test set of 200 web pages. Among them, 100 pages are labeled as relevant to the survey, and the remaining are labeled as irrelevant. The accuracy is measured by the percentage that the agent correctly predicts a web page as relevant or irrelevant.

Opinion Gathering. We first sampled 300 relevant web pages and manually collect the opinions for evaluation. To calculate the accuracy of opinion gathering, we consider the web pages that have all the opinions collected as the correct predictions.

Attribute Extraction. We manually labeled 300 opinions, regarding the gender, language, and required categories from the corresponding survey questions. Accuracy is calculated by computing the number of opinions having all attributes correctly extracted.

D.4 Search Query Generation Analysis

We show the differences in the generated search queries of different LLMs in Table 8. All of these models are used as backbone models for the LLM agent in the genetic algorithm process. Results show that as the model size progressively smaller, we observe less diverse sets of search queries, resulting in less relevant web pages and opinions and overall - lower correlation with actual surveys, as shown in Table 1 of Section 3.5.1. LLAMA3.3-INSTRUCT 70B generates the most diverse set of search queries, covering all of the categories in the example survey question, hence we use this model as the default model of SURVEYPILOT.

E Survey Dataset Information

To perform extrinsic and intrinsic evaluation in our experiments, we have collected 40 questions and findings released in 2024 from established surveys, mostly from the PEW Research Center. Table 9, Table 10, Table 11, and Table 12 show the information of survey questions in each domain, including technology, entertainment, religion, and cuisine, respectively. The available details are the content of the questions, country and gender availability, and the sources of the questions.

In each table, we denote the "Gender" column as the availability of gender information in the surveys (Yes or No). In total, there are 10 questions with gender information of participants available. Regarding the "Country" column, the questions either for US or multiple-country (i.e. US, Italy, Germany, France, and Spain) participants. There are 17 survey questions that have participants from multiple countries.

F Model Deployment

To deploy LLMs for inference, we use two TESLA-A100-80GB GPUs for LLAMA3.3-INSTRUCT 70B and QWEN2.5-INSTRUCT 32B, other models use 1 TESLA-A100-80GB GPU. Models are served in bfloat16 and with the vLLM library (Kwon et al., 2023).

Example Survey Question
How often do you leave a tip in the following service settings?" The response options could include: - Eating at a restaurant where there are servers (e.g., a sit-down restaurant) - Getting a haircut - Having food delivered (e.g., through a delivery app) - Buying a drink at a bar - Using a taxi or rideshare service - Buying a coffee or other beverage at a coffee shop - Eating at a restaurant where there are no servers (e.g., a fast-casual restaurant)

Model	Generated Queries	# Pages	# Opinions
LLAMA3.3-INSTRUCT 70B	- how often do you tip restaurant servers Reddit X - tipping in restaurants Reddit and X - discussions about restaurant tipping frequency Reddit / X - haircut tipping frequency Reddit / X - how often do you tip barbers Reddit X - tipping hairdressers opinions Reddit X - food delivery tipping habits Reddit / X - how often do you tip delivery apps Reddit and X - discussions about tipping delivery drivers Reddit / X - buying drinks at a bar tipping Reddit / X - how often do you tip bartenders Reddit / X - tipping at bars Reddit / X - tipping in taxis or rideshare services Reddit X - how often do you tip Uber or Lyft drivers Reddit and X - discussions about tipping ride-sharing Reddit / X - buying coffee tipping Reddit and X - how often do you tip baristas Reddit / X - tipping at coffee shops opinions Reddit / X - no server restaurants tipping Reddit / X - how often do you tip at fast-casual places Reddit / X - discussions about tipping in fast-casual restaurants Reddit / X	172	20,089
QWEN2.5-INSTRUCT 32B	- how often do you tip at a sit-down restaurant with servers on Reddit or X - opinions on tipping hairdressers or barbers on Reddit or X - discussions about tipping food delivery services on Reddit or X - views about tipping at a bar when buying drinks on Reddit or X - how often do you tip taxi or rideshare drivers on Reddit or X - opinions on tipping at a coffee shop on Reddit or X - how often do you tip at a fast-casual restaurant without servers on Reddit or X	101	12,138
QWEN2.5-INSTRUCT 14B	- opinions on tipping in restaurants with servers on Reddit / X - views about tipping hairdressers on Reddit / X - opinions on tipping food delivery apps on Reddit / X - views about tipping bar staff on Reddit / X - opinions on tipping taxi drivers - views about tipping coffee shop baristas on Reddit / X - opinions on tipping in no-server restaurants on Reddit	31	3,864
LLAMA3.1-INSTRUCT 8B	- How often do people tip in different service settings on Reddit / X - Reddit / X discussions on tipping habits for various services - Do you tip in these service scenarios: restaurant, haircut, delivery, etc.? Reddit	25	1,113

Table 8: Search query comparison between different LLMs for an example survey question.

Question	Source	Gender	Country
Which of the following examples involve the use of artificial intelligence (AI)? Respondents could select multiple choices of the following options: - Wearable fitness trackers that analyze exercise and sleeping patterns - A chatbot that immediately answers customer questions - Product recommendations based on previous purchases - A security camera that sends an alert when there is an unrecognized person at the door - A music playlist recommendation - The email service categorizing an email as spam	PEW - Public Awareness of Artificial Intelligence in Everyday Activities	No	US
How does the increased use of artificial intelligence (AI) in daily life make you feel? Respondents could select one of the following options: - More concerned than excited - More excited than concerned - Equally concerned and excited	PEW - Public Awareness of Artificial Intelligence in Everyday Activities	No	US
How frequent do you interact with artificial intelligence (AI? The response options could include: Almost constantly / Several times a day About once a day / Several times a week Less often	PEW - Public Awareness of Artificial Intelligence in Everyday Activities	**Yes**	US
Do you believe the U.S. government is sharing all it knows about drones with the public? The response options could include: - Telling the public all it knows - Keeping information from the public	CBS News poll: Who's behind the drones?	No	US
Who do you believe is controlling the drones? The response options could include: - U.S. Government - Private Citizens - Foreign Country - Aliens - Don't Know	CBS News poll: Who's behind the drones?	No	US
Do you believe drones are a threat to the U.S.? The response options could include: - Yes, a threat - No, not a threat	CBS News poll: Who's behind the drones?	No	US
Why Data Privacy Is Important? The response options could include: - Avoid Data Breaches - Protect Personal Information - Safeguard Sensitive Data - Comply with Privacy Regulations - Protect Against Surveillance - Maintain Confidentiality - Build Trust with Customers - Preserve Individual Rights - Prevent Identity Theft - Prevent Unauthorized Access	PrivacyEngine - Data Privacy Statistics Worldwide for 2024	**Yes**	**Multiple**
Data Privacy Best Practices The response options could include: - Strong and unique passwords - Updated on data breaches/ security - Phishing/scam email vigiliance - Two-factor authentication - Social media privacy setting reviews - Encrypting sensitive data - Using virtual private network (VPN) on public Wi-Fi - Regular back-ups - Regular software updates - Limit personal sharing online	PrivacyEngine - Data Privacy Statistics Worldwide for 2024	**Yes**	**Multiple**
How do you usually keep track of your passwords for online accounts? Respondents could select one or more of the following options: - Write passwords down - Save passwords in their browser - Frequently reset passwords	PEW - How Americans View Data Privacy	No	US
How do you think companies using AI to collect and analyze personal information will use the information? Respondents could select one or more of the following options: - The information will be used in ways that people would not be comfortable with. - The information will be used in ways that were not originally intended. - The information could make people's lives easier.	PEW - How Americans View Data Privacy	No	US

Table 9: Survey questions in the Technology domain.

Question	Source	Gender	Country
How many hours do you spend playing video games per week?" Response options could include: - Less than 1 hour - 1 to 5 hours - 6 to 10 hours - 11 to 15 hours - 16 to 20 hours - More than 20 hours - I don't play video games - Don't know	Simulation Games Statistics 2024	No	US
Which of the following film genres would you like to see offered more frequently in cinemas? (Please select all that apply.) - ActioNodventure - Comedy - Drama - Romance - Thriller - Sci-Fi - Horror - Documentary - Family/Animation - Other (please specify)	Global Cinema Federation Movie-Goer survey	**Yes**	**Multiple**
Which of the following video game genres do you play or enjoy? (Select all that apply.) - Shooter - Action Adventure - Sports - Multiplayer Online Battle Arena (M.O.B.A.) - Racing - Puzzle - Simulation - Strategy - Action Platform - Battle Royale - Fighting - Role Playing (R.P.G.) - Massively Multiplayer Online (M.M.O.) - Party Games	Statista Report 2024	**Yes**	**Multiple**
Which devices do you primarily use for gaming? Respondents could select all that apply from the following options: - Mobile - Console - PC/Laptop - Tablet - VR Headset	Statista Report 2024	**Yes**	**Multiple**
Which of the following media channels do you use to access news content? (Select all that apply.) - Online channels - Television (broadcast or cable) - Social media (including messaging apps) - Broadcast radio - Physical print media	Digital 2024 Global News Report	**Yes**	**Multiple**
What social media platforms do you prefer to read content from? (Select all that apply.) - Facebook - Instagram - Whatsapp - TikTok - X / Twitter	Digital 2024 Global News Report	**Yes**	**Multiple**
How do you typically listen to music, and what percentage of your total listening time do you spend on each of the following methods? (Please ensure your percentages sum to 100 - Audio Streaming (e.g., Spotify, Apple Music) - Video Streaming (e.g., YouTube, TikTok) - Music on the Radio (broadcast radio, internet radio stations) - Purchased Music (CDs, vinyl, DVDs, digital downloads) - Other Forms of Music Listening (TV, on-demand premium video services, etc.) - Live Music (including livestreams)	Statista Report 2024	No	**Multiple**
Which of the following music genres do you listen to most frequently? (Select all that apply.) - Pop - Rock - Dance/Electronic/House - Soundtracks - Hip-Hop/Rap/Trap - Singer/Songwriter - Classical/Opera - R&B - Soul/Blues - Metal	Statista Report 2024	No	**Multiple**

Table 10: Survey questions in the Entertainment domain.

Question	Source	Gender	Country
How do you regularly participate in religious services? Respondents could select one of the following options: - Only attend religious services in person - Only watch religious services online or on TV - Attend in person AND watch online or on TV - Neither attend in person nor watch online or on TV	PEW - Why some Americans prefer to go to religious services in person and others prefer to watch virtually	No	US
How many of your friends share the same religion as you? Response options could include: All Most Some Hardly any None	PEW - A majority of Americans have a friend of a different religion	No	US
How much discrimination do you think exists against the Jews in our society today? Response options could include: - A lot - Some - Not much - None at all	PEW & EU Survey of Immigrants and Descendants of Immigrants	**Yes**	**Multiple**
How important is it to you that a leader of your country has the following qualities? (Please rate each statement on a scale of 1 to 5, where 1 = Not at all important and 5 = Very important.) - Stands up for people who share your religious beliefs. - Has strong religious beliefs, even if they are different from your own. - Has religious beliefs that are the same as your own.	PEW - Many around the globe say it's important their leader stands up for people's religious beliefs	No	**Multiple**
How important do you think being a member of your country's predominant religion is for being truly [insert nationality of respondent]? (Please select one option.) - Not at all important - Not very important - Somewhat important - Very important	PEW - Views on the importance of religion to national identity	No	**Multiple**
How much discrimination do you think exists against the Jews in our society today? Response options could include: - A lot - Some - Not much - None at all	PEW & EU Survey of Immigrants and Descendants of Immigrants	**Yes**	**Multiple**
Since the start of the Israel-Hamas war, do you feel that discrimination against Jews has increased?" Response options: - Yes, it has increased - No, it has not increased - Not sure	PEW - Rising Numbers of Americans Say Jews and Muslims Face a Lot of Discrimination	No	US
Since the start of the Israel-Hamas war, do you feel that discrimination against Muslims has increased?" Response options: - Yes, it has increased - No, it has not increased - Not sure	PEW - Rising Numbers of Americans Say Jews and Muslims Face a Lot of Discrimination	No	US
As a Jewish, have you ever felt offended by something you saw on the news or social media related to the Israel-Hamas war? Response options: "Yes" or "No."	PEW - Rising Numbers of Americans Say Jews and Muslims Face a Lot of Discrimination	No	US
As a Muslim, have you ever felt offended by something you saw on the news or social media related to the Israel-Hamas war? Response options: "Yes" or "No."	PEW - Rising Numbers of Americans Say Jews and Muslims Face a Lot of Discrimination	No	US
As a Jewish, have you ever stopped talking to someone in person—or unfollowed/blocked someone online—because of something they said about the Israel-Hamas war? Response options: "Yes" or "No."	PEW - Rising Numbers of Americans Say Jews and Muslims Face a Lot of Discrimination	No	US
As a Muslim, have you ever stopped talking to someone in person—or unfollowed/blocked someone online—because of something they said about the Israel-Hamas war? Response options: "Yes" or "No."	PEW - Rising Numbers of Americans Say Jews and Muslims Face a Lot of Discrimination	No	US
Should speech expressing support for 'Israel's right to exist as a Jewish state' be allowed? Answer options: - Allowed - Not allowed - Not sure	PEW - Rising Numbers of Americans Say Jews and Muslims Face a Lot of Discrimination	No	US
Should speech supporting 'Palestinians having their own state' be allowed? Answer options: - Allowed - Not allowed - Not sure	PEW - Rising Numbers of Americans Say Jews and Muslims Face a Lot of Discrimination	No	US

Table 11: Survey questions in the Religion domain.

Question	Source	Gender	Country
How often do you leave a tip in the following service settings? The response options could include: - Eating at a restaurant where there are servers (e.g., a sit-down restaurant) - Getting a haircut - Having food delivered (e.g., through a delivery app) - Buying a drink at a bar - Using a taxi or rideshare service - Buying a coffee or other beverage at a coffee shop - Eating at a restaurant where there are no servers (e.g., a fast-casual restaurant)	PEW - Tipping Culture in America: Public Sees a Changed Landscape	No	US
Compared to five years ago, how do you feel about the number of places where tipping is expected? The response options could include: - Tipping is expected in more places. - Tipping is expected in about the same number of places. - Tipping is expected in fewer places.	PEW - Tipping Culture in America: Public Sees a Changed Landscape	No	US
Which three cuisines do you enjoy the most? (Select up to three) Options: - British - Italian - Chinese - Indian - American - Japanese - French - German - Other (please specify)	YouGov EuroTrack Survey	No	**Multiple**
Which one cuisine is your ultimate favourite? Options: (Same as above) - British - Italian - Chinese - Indian - American - Japanese - French - German - Other (please specify)	YouGov EuroTrack Survey	No	**Multiple**
Which one cuisine do you consider the worst? Options: (Same as above) - British - Italian - Chinese - Indian - American - Japanese - French - German - Other (please specify)	YouGov EuroTrack Survey	No	**Multiple**
Which three cuisines do you enjoy the least? (Select up to three) Options: (Same as above) - British - Italian - Chinese - Indian - American - Japanese - French - German - Other (please specify)	YouGov EuroTrack Survey	No	**Multiple**
How do you prefer your sandwiches to be served? The options could include: - Cut into rectangles - Cut into triangles - Whole and uncut	New York Post's Survey on America's Favorite Sandwiches (October 2024)	No	US
What do you consider the most critical components to making the perfect sandwich? The options could include: - High-quality meat - Airy bread - Using all the right condiments - Fresh veggies	New York Post's Survey on America's Favorite Sandwiches (October 2024)	No	US

Table 12: Survey questions in the Cuisine domain.

```
Imagine you are a {gender} from {country}.

Answer the following question from this perspective.
Others will read what you choose; your goal is to convince them it was chosen from the perspective of
 the persona described above.
Select exactly one option. Do not include any extra commentary.

Answer by typing the number(s) corresponding to your chosen answer(s).

Question: {survey_question}
Options: {numbered_options}
```

Listing 3: Opinion synthesis prompt template.

```
Given the user requirements:
{user_requirements}

The user requirements have outlined some answer options to the question. List them here for me.

Please remember to:
- Provide the options in a single list like this: [\"option 1\", \"option 2\",...]
- Only response with the list of options, do not response with anything else.
- The option names should be the same as the ones provided in the user requirements.
```

Listing 4: Answer option extraction prompt template in Formatting Survey Question

```
Given the user requirements:
{user_requirements}

The user requirements have outlined some answer options to the question. Come up with some
alternatives for the following answer option:
{answer_option}

Please remember to:
- Provide the alternatives in a single list like this: [\"alternative 1\", \"alternative 2\",...]
- Only response with the list of alternatives, do not response with anything else.
- You can come up with alternatives by paraphrasing the terms in the answer option. Do not change the
 original meaning significantly.
```

Listing 5: answer option modification prompt template in Formatting Survey Question

```
Given the user requirements:
{user_requirements}

And the following title and content of a web page:
- Title: {title}
- Content: {markdown_content}

Is the web page relevant to collect human opinions based on the user requirements? Say Yes or No.

Please remember to think about the similarity between the page content and the user requirements
before you answer.

Your answer:
```

Listing 6: Web page filtering prompt template.

```
Come up with some search queries that can be used by Google Search to find human opinions that are
relevant to the user requirements. Your goal is to avoid leading language, represent all sides fairly
, and ensure balanced coverage of options.

Here is the user requirements:
{user_requirements}

Please remember to:
1. Provide the queries in a single list like this: [\"query 1\", \"query 2\",...]
2. Only response with the list of search queries, do not response with anything else.
3. Neutrality:
- Avoid leading terms (e.g., "support," "oppose," "hate").
- Do not assume a default stance (e.g., "Why is [Candidate X] bad?").
4. Cover All Options:
- Explicitly include all major candidates, categories, parties, or viewpoints (e.g., Trump, Biden,
third-party candidates for election survey).
- You may paraphrase the options to ensure better coverage.
5. Balance Perspectives:
- Generate queries that explore both positive and negative sentiments for each option (e.g., "Reasons
 voters criticize [Candidate X]" and "Reasons voters praise [Candidate X]").
6. Open-Ended Exploration:
- Use terms like "opinions on," "discussions about," or "views about" to encourage diverse responses.
7. Avoid Demographics Assumptions:
- Do not assume a specific group (e.g., "young people") unless explicitly required.
8. Simulate an User:
- You may include queries that are question that open a discussion - which are more likely to find
discussions.
9. Language use:
- If the question includes nationality of the survey attendance or any required nationality to answer
 the question, please come up with native queries in that corresponding language. Otherwise, use
English.
```

Listing 7: Search query generation prompt template.

```
Given the user requirements:
{user_requirements}

And the following human opinion collected from {source}:
- Opinion: {content}

Identify the language of the opinion.

You should response in the following format by filling in the placeholders below:
```
[
    "your_thoughts": "your thoughts on the opinion",
    "language": "the language of the opinion",
    "gender": "the gender of the author (if identifiable)"
]
```

Please remember that:
- You have to provide your thoughts on the opinion before identifying the language and gender.
- If the language or gender cannot be identified, put "null" in the corresponding placeholder.
```

Listing 8: Demographic extraction prompt template

````
Given the user requirements:
{user_requirements}

And the following human opinion collected from {source}:
1. The {source} post:
"{post_content}"

2. The opinion to the post:
"{content}"

Here is the list of answer options for your reference - you can only use the names in this list:
{answer_options}

Following the user requirements, the opinion may give an answer(s) to the categorical question.
Identify the answer options mentioned or implied in the opinion. Firstly, provide your thoughts on
the opinion - does the opinion mention or imply any options?

You should response in the following format by filling in the placeholders below:
```
[
    "your_thoughts": "your thoughts on the opinion",
    "identifiable": "true / false",
    "answer_options": "list of option names that the author has interest in, separated by comma"
]
```

Please remember that:
- You have to provide your thoughts on the opinion before identifying the answer options.
- Focus on the opinion only, the post context is just there to improve your understanding.
- Opinions can mention or imply multiple answer options of preferences.
- If the answer options can be identified or implied, put "true" in the "identifiable" placeholder
and put the lists of categorical names in the corresponding placeholders.
- If the answer options cannot be identified or implied, put "false" in the "identifiable"
placeholder and put "null" in the "answer options" placeholders.
````

Listing 9: Answer option extraction prompt template

````
Given the user requirements:
{user_requirements}

And the following HTML content of a web page:
{html_content}

Here are the last {n_opinion} opinions that you have gathered in your last response:
{last_n_opinions}

Collect all the opinions that appear in the current page. Response with a list of opinion, where each
 element is a JSON object having the following keys:

```
{
    "author": "the username of the author",
    "content": "the content of the opinion",
    "date": "the date when the opinion is posted"
}
```

Finally, after you have responsed with the list of opinions, decide whether there are still opinion
left on the page by responding with one of the following flags:
- [Scroll]: this flag triggers the browser to scroll the content for you, so you can collect more
opinions.
- [Terminate]: this flag terminates the browser, meaning that you have reached the end of the page
and there are no more opinions to collect.
````

Listing 10: Opinion gathering prompt template.

```
Given the following survey question:
{survey_question}

Collect all human opinions on the given web page to answer the survey question.

Please remember that:

- The collected human opinions should be provided in the following JSON format:
{
    "author": "the username of the author",
    "content": "the content of the opinion",
    "date": "the date of the opinion",
    "your_thoughts": "your thoughts on the opinion",
    "language": "the language of the opinion",
    "gender": "the gender of the author (if identifiable)",
    "identifiable": "True / False",
    "answer_options": "list of answer option names in the survey questions that the author has
interest in, separated by comma",
}
- Opinions can mention or imply multiple answer options of preferences.
- If the answer options can be identified or implied, put "true" in the "identifiable" placeholder
and put the lists of categorical names in the corresponding placeholders.
- If the answer options cannot be identified or implied, put "false" in the "identifiable"
placeholder and put "null" in the "answer_options" placeholders.
- Each page contains a main opinion and multiple comments / opinions below, please collect all
opinions for each page by scrolling down until the end of the page.
```

Listing 11: Opinion collection & processing prompt template for ScrapeGraphAI.

```
Given the following survey question:
{survey_question}

You need to collect human opinions on social media (Reddit and X/Twitter) to answer the survey
question.

Here is an elaborate plan on how you can execute the task:
1. You need to generate some search queries that can be used to find relevant web pages of
discussions to the survey question.
2. Use the provided Google tool to collect web pages and keep the relevant ones.
3. Use the provided PlayWright tool to browse web pages and collect human opinions.
4. Extract the language, gender, and the answer options specified in the survey question (if
identifiable).
- The collected human opinions should be provided in the following JSON format:
{
    "author": "the username of the author",
    "content": "the content of the opinion",
    "date": "the date of the opinion",
    "your_thoughts": "your thoughts on the opinion",
    "language": "the language of the opinion",
    "gender": "the gender of the author (if identifiable)",
    "identifiable": "True / False",
    "answer_options": "list of answer option names in the survey questions that the author has
interest in, separated by comma",
}
- Opinions can mention or imply multiple answer options of preferences.
- If the answer options can be identified or implied, put "true" in the "identifiable" placeholder
and put the lists of categorical names in the corresponding placeholders.
- If the answer options cannot be identified or implied, put "false" in the "identifiable"
placeholder and put "null" in the "answer_options" placeholders.
- Each page contains a main opinion and multiple comments / opinions below, please collect all
opinions for each page by scrolling down until the end of the page.
```

Listing 12: User prompt template for AutoGen.

```
### Manager Agent
Role:
You are the Manager Agent, the orchestrator of the workflow. Your job is to choose which agent should
 act next based on the overall plan and current progress informed by the Planner Agent.

Responsibilities:
- Direct agents (Planner, Executor, Reviewer) on what task to execute next.
- Keep track of progress and adjust the workflow as necessary.

### Planner Agent
You are the Planner Agent. Your responsibility is to decide whether each step of the plan has been
executed successfully and to inform other agents about the next tasks.

Responsibilities:
- Review outputs from the Executor Agent and confirm that the current step in the plan is complete.
- Provide clear instructions regarding the next step.
- Reiterate the overall plan if needed to ensure all agents are aligned.

### Executor Agent
Role:
You are the Executor Agent. Your task is to perform the actual work based on the current step
dictated by the Planner Agent. This includes generating search queries, using web tools to browse and
 collect relevant human opinions, and extracting the required details (language, gender, categories,
etc.) from the content.

Responsibilities:
- Follow the current step's instructions provided by the Planner Agent precisely.
- Utilize available tools (e.g., search tools, browser automation) to gather and extract the required
 data.

### Reviewer Agent
Role:
You are the Reviewer Agent. Your role is to review the outputs from the Executor Agent and provide
feedback on the quality and completeness of the results. You check for format accuracy, diversity in
search queries, and overall adherence to the survey collection plan.

Responsibilities:
- Critically review the Executor Agent's outputs for correctness and adherence to the required
structure.
- Identify any errors in the response format.
- Provide constructive feedback or recommendations to improve the current task if needed.
- Confirm that each tool call aligns with the overall plan.
```

Listing 13: System prompt template for AutoGen.

```
### Crossover
Given the user requirements:
{user_requirements}

And the following 2 sets of search queries for collecting opinions:
- Solution 1: {parent_1}
- Solution 2: {parent_2}

Combine and improve these solutions. Please remember to provide the queries in a single list like
this: [\"query 1\", \"query 2\",...]

### Mutation
Given the user requirements:
{user_requirements}

And the following set of search queries for collecting opinions:
{solution}

Come up with multiple versions of the search queries by substituting some terms with semantically
related alternatives. Please remember to provide the queries in a single list like this: [\"query
1\", \"query 2\",...]
```

Listing 14: Crossover and Mutation prompt template for the proposed Genetic Algorithm.

```
### Reviewer
- System Prompt:
You are a critic analyzing solutions.

- Message:
Given this user requirements:
{user_requirements}

Analyze this solution:
{solution}

Provide constructive feedback.

### Executor
- System Prompt:
You are an author improving solutions.

- Message:
Given this user requirements:
{user_requirements}

And the feedback: {critique}

Improve this solution: {solution}
```

Listing 15: Search queries refinement prompt template for the proposed Genetic Algorithm.