Reimagining Social Research: An Interactive R Toolkit and Guideline for Generating and Validating LLM-Based Social Avatars

Presentation Preference: Parallel Track Session
Track: Social applications of Large Language Models

Shi Wang and Yusung Su

Department of Political Science, Tsinghua University, Beijing, China

1 Introduction

The advent of Large Language Models (LLMs) marks a significant milestone in social computing. Existing literature suggests that LLMs can process and generate human-like text, opening new avenues for simulating social dynamics and conducting virtual surveys [1, 2]. However, most current research remains theoretical. There is an urgent need for practical guidelines and accessible toolkits that empower social scientists to reliably transition from traditional survey data to AI-driven virtual environments. This ongoing study addresses this gap by developing an open-source, UI-integrated R package. More than just a software tool, this package introduces a standardized methodological guideline for researchers to generate, interact with, and empirically validate LLM-based social avatars.

2 System Architecture and Workflow

To operationalize the creation of social avatars, we are developing a user-friendly R package equipped with a graphical user interface (UI), explicitly tailored for social scientists lacking deep computational backgrounds. The workflow is designed as a continuous, intuitive pipeline. Initially, during the data ingestion and contextualization phase, researchers upload existing survey datasets alongside their codebooks. The system parses variables, scales, and contextual metadata, translating quantitative data points into comprehensive qualitative profiles.

Following this, the avatar generation module leverages advanced LLMs via API integration to digest the ingested data. It automatically constructs detailed, prompt-ready character biographies for each respondent, effectively creating a digital twin imbued with specific socio-demographic backgrounds and values. Finally, the interactive querying interface allows researchers to interrogate these generated avatars on novel or hypothetical social issues. Crucially, the toolkit goes beyond deterministic responses by supporting distributional outputs—running multiple simulated iterations to gauge response probability

and stability. It also enables researchers to isolate specific individuals (e.g., "Avatar #1042") for virtual follow-up qualitative interviews, blending macro-level survey analysis with micro-level narrative exploration.

3 Validation Methodology: The CGSS 80/20 Split

A paramount challenge in LLM-based simulation is mitigating "hallucinations" and ensuring construct validity. To establish this toolkit as a scientifically rigorous guideline, we introduce a quantifiable validation protocol using data from the Chinese General Social Survey (CGSS). The CGSS provides a rich, multi-dimensional repository of socio-demographic and attitudinal data, making it an ideal bedrock for testing avatar fidelity.

Our validation procedure adapts a machine-learning-inspired 80/20 split approach applied directly to individual survey items. For a given CGSS respondent, 80% of their survey responses (encompassing core demographics, baseline values, and behavioral traits) are utilized to prompt and contextualize the LLM, establishing a robust baseline character biography. In the subsequent prediction task, the resulting avatar is prompted to answer the remaining 20% of the survey questions (such as nuanced attitudes toward current events, social trust, or economic outlooks) which the model has not previously encountered. To benchmark the performance, these AI-generated responses are quantitatively compared against the actual human responses—the ground truth—extracted from the CGSS dataset. By measuring predictive accuracy, response variance, and attitudinal alignment, we can systematically evaluate the reliability of the LLM avatars.

4 Ethical Considerations and Expected Contributions

As this toolkit aims to establish a methodological standard, it embeds ethical constraints directly into its architecture. We establish programmatic guidelines for ensuring the strict anonymization of participant data before API transmission, thereby protecting privacy [1]. Furthermore, the package includes audit functions to check for algorithmic bias by comparing the distribution of avatar responses against original, real-world survey distributions.

Ultimately, this ongoing study will contribute to the ICSC 2026 community in three distinct ways. First, it delivers a tangible, open-source R toolkit that democratizes LLM-avatar generation for the broader social science community. Second, it introduces a rigorous 80/20 validation framework using CGSS data to empirically quantify the reliability of AI simulations. Finally, it proposes a practical, ethically grounded guideline for conducting the next generation of AI-augmented computational social research.