How to Use AI in Experimental and Survey Research: A Practical Guide for Researchers of the WULab

Artificial intelligence, particularly large language models (LLMs), is rapidly transforming experimental and survey research. When used thoughtfully, these tools can streamline questionnaire development, enable scalable pretesting, and simulate human behavior for theory testing. This guide summarizes current best practices, opportunities, and pitfalls to help researchers in our lab apply AI responsibly and effectively in their own studies.

AI-Assisted development of Experimental Instructions and Surveys

1. Drafting

- Provide the LLM with your research question, target population, and desired number of items. The model retrieves relevant examples (e.g., from survey databases like GESIS SQP) and recent public discussions to create draft questions (Adhikari, Hartland, Weber, and Cannanure, 2025).
- Let the LLM draft example instructions for standardized economic games and use them as a starting point.
- Use LLMs to tailor instructions and survey items for a specific subsample (Adhikari, Hartland, Weber, and Cannanure, 2025). For example, check for jargon, detect cultural mismatches or adapt language for prticular audiences such as children or respondents from other countries.
- Utilize LLMs to simplify and clarify your instructions by identifying redundancies, suggesting clearer rephrasing, and shortening text without losing meaning. This makes instructions more concise and easier to follow.

2. AI Assisted Pilot Testing and Refinement

- Run an LLM-simulated pilot: Use the LLM to play an interviewer and participants with distinct personas (short descriptions of imagined participants). The AI participants respond to your draft items, and the interviewer asks clarifying follow-ups. This mimics a cognitive interview, revealing unclear wording or hidden bias before you field the survey with humans. This can be particularly helpful for non-standard target samples.
- Automated expert review: Feed the pilot transcripts to another LLM acting as a reviewer. It produces a structured table showing which questions caused problems, what kind of issues occurred (e.g., ambiguity, bias), and which persona was affected. This provides a transparent, evidence-based basis for revision.
- Test your experiments and surveys technically with LLMS to identify potential programming errors and debug experimental flow. Helpful tools include:

    - Botex: python package to run LLM as otree participant (https://github.com/trr266/botex).
    - Alter ego: python library to run experiments with LLMs (https://github.com/mrpg/ego).

AI for Simulating Human Behavior

- Use AI agents to simulate human behavior in experiments or surveys to approximate participant responses (Sarstedt et al. 2024). Advanced methods, such as conditioning LLMs on human interview data (Park et al., 2024) or demographic backstories (Argyle et al., 2023), can reach high accuracy. Consider the needed complexity for your study before implementation.
- Mock datasets can assist with:
    - Estimating effect sizes for power analysis.
    - Preparing analysis scripts, for example, for preregistered reports
- Challenges in AI simulations (see e.g., Anthis et. al. 2025)
    - Be cautious: Seemingly small technical choices (e.g., prompts, temperature) can drastically affect outcomes
    - Diversity and bias: AI outputs often sound too similar and reflect their training data, which can overrepresent certain worldviews. Counteract this by including balanced demographic prompts or human review.
    - Sycophancy: Models tend to agree with the researcher's framing—avoid leading prompts and instruct models to predict human-like responses honestly rather than aiming to please.
    - Alienness: Though AI outputs sound human-like, their reasoning differs. Always validate simulations against known human data.
    - Generalization: AI simulations may work for one dataset but fail others. Test models across multiple populations and preregister analytic decisions.

AI in Experiments

- You can use AI bots as "dummy" players in experiments when you are interested only one side's reaction in an interaction.
- Including chatbots in your experiments is also possible.
    - You can use existing oTree plugins (https://github.com/clintmckenna/oTree_gpt)
    - You can develop your own chatbot using an LLM API (at WU: via Azure)
- The LUCID Framework (2025) introduces the concept of using LLMs as 'confederates'—controlled conversation partners in experiments. With the LUCID Toolkit researchers can integrate consistent AI agents into Qualtrics studies.
- Mixed-subject-designs (Broska, Howes, and van Loon, 2025) combine AI-generated data with human data. AI predictions act as informative but imperfect observations, allowing comparison with human responses to test psychological theories and the reliability of AI-based findings.

Ethical and Methodological Recommendations

- Always disclose if and how AI was used, including model type, prompt wording, and data sources.
- When real participant data inform AI models, obtain proper consent.
- Treat AI-generated participants as distinct from human participants.
- Validate all results empirically before publication.

References

Adhikari, D. M., Hartland, A., Weber, I., & Cannanure, V. K. (2025, July). Exploring LLMs for automated generation and adaptation of questionnaires. *In Proceedings of the 7th ACM Conference on Conversational User Interfaces* (pp. 1–23). https://dl.acm.org/doi/10.1145/3719160.3736606

Anthis, Jacy Reese, et al. "Llm social simulations are a promising research method." *arXiv preprint arXiv:2504.02234* (2025).

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Broska, D., Howes, M., & van Loon, A. (2025). The mixed subjects design: Treating large language models as potentially informative observations. *Sociological Methods & Research.* https://doi.org/10.1177/00491241251326865

Cummins, J. (2025). The threat of analytic flexibility in using large language models to simulate human data: A call to attention. arXiv preprint arXiv:2509.13397

LUCID Framework. (2025). LLM-Unified Confederate for Interactive Dialogue. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5256150

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

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing, 41(*6), 1254–1270. https://doi.org/10.1002/mar.21982

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22(*11), 1359–1366. https://doi.org/10.1177/0956797611417632