Optimizing Behavioral Experiment Design with In-Silico Simulations

Sanchaita Hazra[α,β] **Pao Siangliulue**[β] **Bodhisattwa Prasad Majumder**[β]
Peter Clark[β] **Amy X Zhang**[γ,β] **Joseph Chee Chang**[β]

[α]University of Utah, Salt Lake City, UT
[β]Allen Institute for AI, Seattle, WA
[γ]University of Washington, Seattle, WA

Email: sanchaita.hazra@utah.edu

Abstract

Running behavioral experiments is expensive, slow, and vulnerable to design flaws detectable only after data collection. We introduce a preliminary multi-agent framework that uses simulations to iteratively refine behavioral study designs before recruiting human participants. With personas instantiated through target demographic attributes, we elicit explanatory reasoning traces from LLM simulations that serve as diagnostic evidence of structural flaws. In two survey case studies, simulation-rich Critic substantially outperforms a design-only baseline on the precision of high-severity structural errors, while providing little value for identifying missed opportunities. These early results suggest reasoning traces can expose a qualitatively distinct class of design flaws inaccessible to the design-only baseline.

1 Introduction

Behavioral experiments have produced scientific knowledge worth the effort of conducting them. Social experiments for Prospect Theory (Kahneman & Tversky, 1979), Endowment Effect (Kahneman et al., 1991), Milgram Obedience (Blass, 1991), Asian disease problem/Framing/Anchoring experiments (Chapman & Johnson, 1999; Nelson et al., 1997; Tversky & Kahneman, 1981) yield evidence of systematic departures from rational human behavior. Survey experiments, e.g., the General Social Surveys (Davis & Smith, 1991), the British Election Study (British Election Study, 1964), and the World Values Survey (World Values Survey Association, n.d.), have been path-breaking in their approach to identifying systematic patterns in human behavior.

While designing and running human-participant experiments are critical, they pose obvious limitations: (1) Conducting experiments is costly and often requires substantial funds. Prior work has reported an average cost per participant per hour of roughly €9.06 (Christensen et al., 2017), (2) Recruitment is time-consuming and difficult (Ritter et al., 2012), and can take up to months or years for field surveys, (3) Faulty design decisions that often slip through, such as missing answer options, hidden confounds, and experimenter demand effects, are often identified only after data are collected (Presser et al., 2004). A poorly designed experiment not only wastes recruitment time and budget but can also lead to incorrect or incomplete research outputs.

The traditional remedy to avoid the costly process of redesigning and rerunning an experiment is study piloting. However, the quality of each revision is often limited by the pilot size and the extent of the information captured through the pilot experiments. Today, large language models (LLMs) present new opportunities to address this challenge. A parallel industry has emerged around low-cost, more diverse "synthetic respondents", where AI-generated personas are used to approximate survey answers (Zhao et al., 2025; González-Bustamante et al., 2025; Manning & Horton, 2026; Cao et al., 2025; Liu et al., 2026)

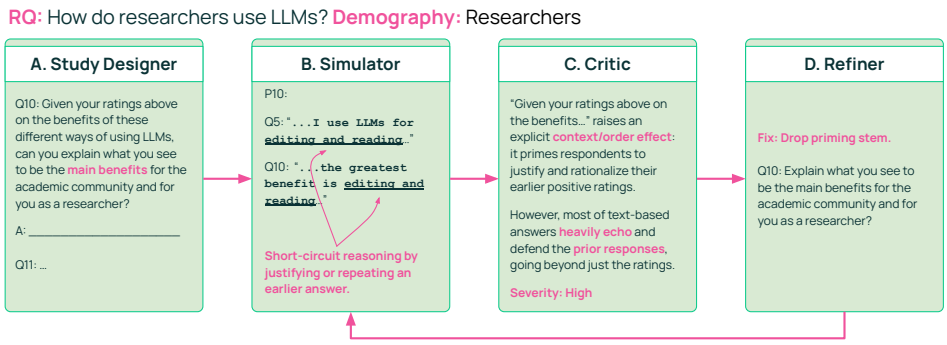

Figure 1: **The iterative loop** for optimizing study design with in-silico simulations. The Critic identifies a behavioral error that surfaced only in **simulated reasoning**, and hence could not be captured solely from the design.

or gauge market behavior (start-ups including Expected Parrot, Simile AI) before recruiting human participants.

In this study, **we aim to leverage simulations to improve study design**, instead of directly revoking the need for human participants[1]. We propose a multi-agent iterative loop framework to critique and improve a study design using responses and reasoning traces from simulated LLM participants. ***We investigate whether reasoning traces from simulations reliably provide enough information to surface structural study design flaws and missed opportunities.*** Our framework builds on the hypothesis that the reasoning traces generated by a simulated participant in answering a study question can be diagnostically useful even when the final answer is not a faithful prediction of any real human participant.[2] In our proposed framework, reasoning traces supersede the accuracy of the simulator.

This paper reports our framework, early explorations, and the following contributions:

1. We stake out a positioning for LLM social simulation as an instrument for design critique, distinct from both higher-fidelity simulation and human replacement, where the value lies in reasoning traces as a diagnostic signal rather than in synthetic answers as data.
2. We describe an early four-component loop (Study Designer, Simulator, Critic, and Study Refiner) that mirrors the human pilot-and-revise workflow and is self-correcting, so that a bad fix can be identified on the next round.
3. We report preliminary, qualitative results from two prototype case studies, framed against a no-simulation baseline that critiques the study design alone, without reasoning traces.

2 Framework

We frame an iterative loop system (Figure 1). Starting from a research question, we sequentially implement the agentic components: Study Designer, Simulator, Critic, and Study Refiner, to converge on an **optimal study design** that best captures a scientist's (the intended user) research intent in an iterative refinement process.

A. Study Designer. The designer converts a user's input, i.e., a research question and optional target demographics, into a first draft of the study design, which entails producing questions and answer options for respondents. For more stimulus-based experiments, e.g., in economics or psychology, the input could also include stimuli. Building on the observation of reasonable draft study generation by LLMs in (Adhikari et al., 2025), we use an LLM to generate the complete study design, which is computationally operable by

[1]For our study, we scope this work to survey experiments, strategic games, and incentivized social science experiments. We explicitly exclude clinical trials in biology, public health, and neuroscience.

[2]A simulated survey study of *how researchers with different demographic backgrounds use LLMs for research* yields many reasoning traces such as "*translation is also a frequent use for me, but translation has nowhere to go in answer options, so it collapses into writing and editing*", exposing an important missing-answer-option error in the study design.

programmed (e.g., LLM) participants. When the demographics is not provided as input, the final study should also include a proposed demographic specification.

B. Simulator. The Simulator instantiates LLM-based personas in the study design, comprising demographic profiles and humanizing characteristics that help them produce aligned reasoning and responses for study questions.

- **Demographic Profiles**: Each profile is algorithmically sampled from the dimensions provided in the study design (e.g. gender, field of study, first language). This unique demography snapshot is deliberately *study-agnostic*, conditioning on only the target demography and never on any study design topic.
- **Humanizing Statements**: This includes statements about the humanizing aspects of a persona, e.g., their background, routines, relationships, and tastes. We use LLM-as-a-judge to check for logical consistency when merging persona statements with sampled demographic profiles, ensuring minimal contradiction (e.g., a 19-year-old with a decade of professional experience is contradictory). Finally, we check for leakage using an LLM by removing any humanizing statements that reproduce a study question's answer option, ensuring that the persona does not pre-encode the responses the study is meant to elicit.
- **Reasoning Traces:** Since our main hypothesis is to check if reasoning traces help improve study design, we instruct the simulated participants to produce long reasoning chains before committing to an answer for a particular question.

C. Critic. Given the research question, the study design, and the reasoning traces, the Critic surfaces two types of outputs: (a) it flags *methodological and structural* design flaws, and (b) it proposes *missed opportunities*, i.e., factors that could explain the study outcome or fulfill the research question better but are currently missing from the design. With simulation, the Critic can examine the responses and reasoning traces from simulated participants to ground its conclusions in evidence. Along with proposed errors and missed opportunities, it also reports the *perceived severity* of each item with three levels: high, medium, and low.[3]

Given that the Critic could take the reasoning traces for all participants for questions in a design at once, there is a risk of context blow-up when the Critic is designed as a single LLM call. We experiment with alternate structures to allow reasoning scaling by providing access to simulation traces for a single question at a time. When the Critic processes the full study, along with all the simulated reasoning traces for all questions at once, we call it *All-Q*. In an ablation named *One-Q*, we provide Critic the research question, all questions without answer options from the study design, and a single question-answer block, along with all participants' responses and reasoning traces for the block.

D. Refiner. The Refiner converts the Critic's assessments into concrete fixes and applies them to the existing study design, producing the input for the next simulation cycle. Often, errors are fixed by editing an existing erroneous question, but addressing many missed opportunities may result in additional questions. We plan to explore adding a budgetary constraint to steer the fixing.

3 Preliminary Results and Discussion

We present preliminary results for the Critic using two case studies involving survey experiments, primarily to establish whether simulation-based reasoning helps expose novel design flaws. For each case study, we simulate 50 LLM participants following the outlined demography in the study design, and generate assessments with (All-Q, One-Q) Critic. We ask an expert rater (co-author for Study 1, Economics PhD student with design familiarity for Study 2) to evaluate if a high-severity structural error/missed opportunity flagged by the Critic is valid. We measure precision, the share of high-severity items as endorsed by an expert-rater, and report them in Table 1.

[3] **High** means the error or missed opportunity, if left unaddressed, could materially compromise the study's conclusions; **Medium** denotes a secondary-priority issue whose resolution would strengthen the robustness of the study's results without being essential to their validity; and **Low** indicates a minor issue of limited relevance that may or may not bear on the primary research question.

	Study 1 (Liao et al., 2024)				Study 2 (Ling et al., 2026)			
	Structural Errors		Missed Opportunity		Structural Errors		Missed Opportunity	
	Precision	High/N	Precision	High/N	Precision	High/N	Precision	High/N
Baseline								
All-Q Critic	0.17	12/55	0	3/11	0	5/42	1	4/9
One-Q Critic	0.00	0/55	0.67	3/11	0.13	8/42	0.75	4/11
With Simulation								
All-Q Critic	**0.46**	13/55	0.25	8/11	**0.6**	5/42	0.8	5/11
One-Q Critic	**0.24**	21/55	0.5	6/13	**0.25**	8/42	0.75	4/14

Table 1: **Precision of high-severity flags** by the Critic, with vs. without simulation.

- **Study 1.** We feed our simulator with the high-level research question, final study design, and target demography of Liao et al. (2024), which conducts a large-scale survey of research article authors on their uses and perceptions of LLMs as research tools.
- **Study 2.** We feed the simulator with the final study design from Ling et al. (2026), which surveys university students' potential social desirability bias using indirect questioning, where students report both their own AI use and that of their peers.

Simulation improves the precision of high-severity structural critiques. Precision increases when simulation traces were provided to surface structural errors; 0.17→0.46, 0→0.24 in Study 1; 0→0.60, 0.12→0.25 in Study 2 (see Table 1), regardless of the Critic choice. Reasoning traces do not merely provide more evidence to validate the already-existing errors but also surface more severe, unique design flaws that were tagged as valid by the experts (see Appendix B)

Critic with all reasoning traces at once performs best. Without simulation traces, All-Q Critic is only stronger in surfacing design flaws in Study 1. Adding reasoning traces from simulations enhances the All-Q Critic to emerge as the strongest. This provides preliminary evidence that isolating reasoning traces for a single question may obfuscate the Critic's ability to spot correlated patterns in reasoning traces across questions. For instance, in Study 1, the All-Q was able to spot a post-hoc justifying phenomenon in simulation reasoning to surface a novel behavioral flaw in design that primed participants to use responses from an earlier question by short-circuiting reasoning for a later, related but different question.

For missed opportunities, simulations add little. We find an opposite pattern for missed opportunities. Missed opportunities evolve directly from the existing study questions in silo, without the need for additional reasoning traces of the simulations; precision for One-Q Critic stands at 0.67 (Study 1) and 0.75 (Study 2). This is partly explained by missed opportunities being directly related to unasked insights that could directly strengthen the evidence for the research question, which may not require understanding the participant's reasoning process.

Thematically, the Critic, with or without simulations, is most successful at identifying structural errors intrinsic to the study questions, for example, *Measurement* and *Behavioral* errors (See Appendix A). Adding simulated reasoning traces does not replace these categories; rather adds a reasoning-dependent validation (See Appendix C). In some cases, adding reasoning traces resurfaces a new error type for the same question (See Appendix D). Another prominent pattern from reasoning traces is how participants reinterpret a study question instrument. Often, participants make a departure from using the suggested instrument in the study question, constructing partially unrelated instruments to guide their answers, alluding to a newer interpretation (See Appendix E).

Open Questions and Conclusion. We show that reasoning produced through a simulation harness is more informative than critiquing a behavioral study design in isolation. The diversity in reasoning plays a crucial role in surfacing new information beyond question semantics that allows the Critic to identify novel flaws and missed opportunities. To amplify diversity, we plan to explore diversity-inducing generation methods. We also observe that the intrinsic reasoning tokens produced by a reasoning model before generating a response possess more behavioral cues than the explanation surfaced as part of the output

tokens. Finally, we plan to demonstrate through user studies that our iterative system, when steered by the researcher-user, will help them to compensate for the LLM's limited domain knowledge but still obtain benefits from the proposed harness.

References

Divya Mani Adhikari, Alexander Hartland, Ingmar Weber, and Vikram Kamath Cannanure. Exploring LLMs for automated generation and adaptation of questionnaires. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*, Waterloo, ON, Canada, 2025. Association for Computing Machinery. doi: 10.1145/3719160.3736606. arXiv:2501.05985.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Position: LLM social simulations are a promising research method. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2504.02234.

Christopher A. Bail. Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024. doi: 10.1073/pnas.2314021121. URL https://doi.org/10.1073/pnas.2314021121.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024. doi: 10.1017/pan.2024.5.

Thomas Blass. Understanding behavior in the milgram obedience experiment: The role of personality, situations, and their interactions. *Journal of personality and social psychology*, 60 (3):398, 1991.

Peter Bohm. Cvm spells responses to hypothetical questions. *Natural resources journal*, pp. 37–50, 1994.

British Election Study. The British Election Study. https://www.britishelectionstudy.com/, 1964. Conducted after every UK general election since 1964; founded by David Butler and Donald E. Stokes, Nuffield College, University of Oxford. Accessed 2026-06-22.

Donald T Campbell and Donald W Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81, 1959.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3141–3154, 2025.

Gretchen B Chapman and Eric J Johnson. Anchoring, activation, and the construction of values. *Organizational behavior and human decision processes*, 79(2):115–153, 1999.

Tina Christensen, Anders H. Riis, Elizabeth E. Hatch, Lauren A. Wise, Marie G. Nielsen, Kenneth J. Rothman, Henrik Toft Sørensen, and Ellen M. Mikkelsen. Costs and efficiency of online and offline recruitment methods: A web-based cohort study. *Journal of Medical Internet Research*, 19(3):e58, 2017. doi: 10.2196/jmir.6716. URL https://www.jmir.org/2017/3/e58/.

Herbert H Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, 12(4):335–359, 1973.

Ronald G Cummings and Laura O Taylor. Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method. *American economic review*, 89(3):649–665, 1999.

James A Davis and Tom W Smith. *The NORC General Social Survey: A User s Guide*. SAGE publications, 1991.

Bastián González-Bustamante, Nando Verelst, and Carla Cisternas. Emulating public opinion: A proof-of-concept of ai-generated synthetic survey responses for the chilean case. *arXiv preprint arXiv:2509.09871*, 2025.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in llms, 2025. URL https://arxiv.org/abs/2509.03730.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. doi: 10.2307/1914185.

Daniel Kahneman, Jack L Knetsch, Richard H Thaler, et al. The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, 5(1):193–206, 1991.

Jon A Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236, 1991.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. Llms as research tools: A large scale survey of researchers' usage and perceptions, 2024. URL https://arxiv.org/abs/2411.05025.

Yier Ling, Alex Kale, and Alex Imas. Underreporting of ai use: The role of social desirability bias. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2026.

Xuan Liu, Haoyang Shang, and Haojian Jin. Cobra: Programming cognitive bias in social agents using classic social science experiments. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–30, 2026.

Benjamin S Manning and John J Horton. General social agents. Technical report, National Bureau of Economic Research, 2026.

Benjamin S. Manning, Kehang Zhu, and John J. Horton. Automated social science: Language models as scientist and subjects. Working Paper 32381, National Bureau of Economic Research, April 2024. URL https://www.nber.org/papers/w32381.

Thomas E Nelson, Zoe M Oxley, and Rosalee A Clawson. Toward a psychology of framing effects. *Political behavior*, 19(3):221–246, 1997.

Martin T Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. In *Sociological methods*, pp. 279–299. Routledge, 2017.

Jinghua Piao, Yuwei Yan, Nian Li, Jun Zhang, and Yong Li. Exploring large language model agents for piloting social experiments. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2508.08678.

Philip M Podsakoff, Scott B MacKenzie, Jeong-Yeon Lee, and Nathan P Podsakoff. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5):879, 2003.

EC Poulton. Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. *Psychological Bulletin*, 91(3):673, 1982.

Stanley Presser, Mick P Couper, Judith T Lessler, Elizabeth Martin, Jean Martin, Jennifer M Rothgeb, and Eleanor Singer. Methods for testing and evaluating survey questions. *Methods for testing and evaluating survey questionnaires*, pp. 1–22, 2004.

Frank E. Ritter, Jonathan H. Morgan, and Jong W. Kim. Practical aspects of running experiments with human participants. In *Proceedings of the 21st Conference on Behavior Representation in Modeling and Simulation (BRIMS)*, pp. 234–240, Amelia Island, FL, 2012. BRIMS Society. URL https://acs.ist.psu.edu/papers/ritterMK12.pdf. Paper no. 12-BRIMS-046.

Gary Saretsky. The oeo pc experiment and the john henry effect. *The Phi Delta Kappan*, 53(9): 579–581, 1972.

Norbert Schwarz, Hans-J Hippler, Brigitte Deutsch, and Fritz Strack. Response scales: Effects of category range on reported behavior and comparative judgments. *Public opinion quarterly*, 49(3):388–395, 1985.

William R Shadish, Thomas D Cook, and Donald T Campbell. Quasi-experiments: interrupted time-series designs. *Experimental and quasi-experimental designs for generalized causal inference*, 1:171–205, 2002.

Roger Tourangeau and Kenneth A Rasinski. Cognitive processes underlying context effects in attitude measurement. *Psychological bulletin*, 103(3):299, 1988.

Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. The psychology of survey response. 2000.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

Stephen J Weber and Thomas D Cook. Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological bulletin*, 77(4): 273, 1972.

World Values Survey Association. World Values Survey. https://www.worldvaluessurvey.org/, n.d. Accessed: 2026-06-26.

Jianpeng Zhao, Chenyu Yuan, Weiming Luo, Haoling Xie, Guangwei Zhang, Steven Jige Quan, Zixuan Yuan, Pengyang Wang, and Denghui Zhang. Large language models as virtual survey respondents: Evaluating sociodemographic response generation. *arXiv preprint arXiv:2509.06337*, 2025.

AI-use Statement

We use LLM-driven coding agents for implementing parts of the framework. All of our simulations (hence, snippets from reasoning) are generated by LLMs, by design. We occasionally used LLMs, including Grammarly, for suggestions on our expositional clarity.

Table 2: Taxonomy of Experiment Design Errors

Cluster	Pitfall	Mechanism
Sample	WEIRD sampling	Structurally unrepresentative sample used as a universal human baseline. (Henrich et al., 2010)
Sample	Sample selection bias	Participation requires access, motivation, or time not everyone has. (Shadish et al., 2002)
Confounding	Confounded manipulation	Multiple features vary across conditions, preventing clean causal attribution. (Shadish et al., 2002)
Confounding	Control condition not inert	The comparison condition has its own effects rather than serving as a neutral baseline. (Shadish et al., 2002)
Confounding	No control group	No comparison point exists to isolate the cause of observed change. (Shadish et al., 2002)
Confounding	Carry-over effects	Prior conditions contaminate later responses in within-subject designs. (Poulton, 1982)
Confounding	Stimuli-as-fixed-effects	Results depend on specific stimuli rather than the broader category. (Clark, 1973)
Confounding	No counterbalancing	Order is confounded with the condition. (Poulton, 1982)
Measurement	Single operationalization	One measure cannot separate the construct from instrument-specific quirks. (Shadish et al., 2002)
Measurement	Response-option coverage	Missing or poorly chosen answer options in the range of response categories. (Schwarz et al., 1985)
Measurement	Common method variance	Shared measurement method inflates observed relationships. (Podsakoff et al., 2003; Campbell & Fiske, 1959)
Measurement	Scale range effects	Scale endpoints anchor responses and define what appears normal. (Schwarz et al., 1985)
Measurement	Hypothetical bias	Stated preferences diverge from revealed behavior. (Bohm, 1994; Cummings & Taylor, 1999)
Measurement	Ceiling/floor effects	The task or scale prevents meaningful variation. (Shadish et al., 2002)
Behavioral	Demand characteristics	Participants infer the hypothesis and adjust their behavior. (Orne, 2017; Weber & Cook, 1972)
Behavioral	John Henry effect	Control-group participants overperform because they know they are being compared. (Saretsky, 1972)
Behavioral	Question order/context effects	Earlier questions prime responses to later items. (Tourangeau & Rasinski, 1988; Tourangeau et al., 2000)
Behavioral	Satisficing	Fatigue or low motivation produces low-effort responses. (Krosnick, 1991)

A Taxonomy of Errors

Referring to prior literature, we have attempted to create a taxonomy of structural errors in experimental study design. We have identified four broad clusters: Sample, Confounding, Measurement, and Behavioral errors. Each cluster also implies a different research function that failed. In Table 2, we identify the pitfalls of the experiment and their mechanism, grouped by each cluster.

B Example: Simulation improves the precision of high-severity structural critiques

For a question in Study 1:
"If future LLMs can prevent hallucinations and can always attribute any copyrighted text to original sources, how ethically acceptable do you find the use of LLMs for: You prompt with an idea and the LLM generates text that you then directly copy and paste into a text editor?"

The critic points to a counterfactual framing error that assumes the two named technical problems are the only barriers to acceptability. Reasoning traces note residual unease, *'even with the technical problems solved... there's a question of authorship,'* indicating that the question premise underspecifies the construct and does not fully address what actually drives their acceptability.

Our expert-rater noted: "Yes. This is consistent with the actual survey. I would call this a missed opportunity; we could have additionally asked 'why' to get the real driver."

C Example: Adding simulated reasoning traces adds a reasoning-dependent validation

Consider a question from Study 2:
"Amongst the students that you know, what percent would you say use LLM(s) in their everyday lives?"

The Critic flags this item as a high-severity measurement error *both* with and without simulation, i.e., the classification remains unchanged when traces are added.

- **Without simulations.** The Critic identifies the flaw from the wording: "students that you know" is an undefined, self-selected reference group that varies across respondents (close friends vs. classmates vs. an entire cohort). Because this item is one side of the own-vs-others comparison at the core of the study, the ambiguity directly threatens the validity of the gap estimate.
- **With simulations.** The same Measurement error is flagged, but the reasoning traces add an additional validation that the question wording cannot reveal independently: each respondent chooses a *different* reference group before answering, so nominally identical responses are not comparable. Three personas anchor on visibly different populations.
  - "Looking around at my cohort *and the undergrads I TA for*, it's nearly universal…I'd estimate …around 85 percent."
  - "Thinking about the students I know — *my labmates, the internationals on my football team, classmates* — almost everyone uses these tools now …around 90 percent."
  - "In *my business graduate cohort*, basically everyone uses these tools …I'd estimate …around 85 percent."

D Example: Adding reasoning traces resurfaces a new error cluster

Question from Study 1: *Given your ratings above on the benefits/usefulness of these different ways of using LLMs, can you explain what you see to be the main benefits/usefulness for the academic community and for you as a researcher?*

- **Without simulations.** The Critic suggested a double-barreled measurement error inherent to these questions as the question attempts to uncover effects on both the respondent and the research community.
- **With simulations.** With traces, the Critic additionally identifies a behavioral error resulting from an order-effect error because the questions begin with *"Given your ratings above,"*. Simulated participants often anchor their explanations to answer options from related, different questions asked earlier. This could hint at the participants trying to post-hoc justify and rationalize the answers to earlier questions instead of thinking through independently for the current questions. Answers heavily echo and defend the prior usefulness ratings.
  - Participant (p002) response in an earlier closed-form multiple choice rating question:
    Q: "How useful/beneficial do you find LLMs to be for: You are writing in a

text editor and the LLM is providing editing support?"
Options: Not at all useful · Slightly useful · Moderately useful · Very useful · Extremely useful · I have not used LLMs for this type of activity
Answer: *Extremely useful*
The same question item was asked for six uses (follows); p002's ratings were: information seeking: *Very useful*, editing: *Extremely useful*, generation: *Very useful*, ideation: *Moderately useful*, data generation: *Slightly useful*, data analysis: *Very useful*)

- A later question asks:
"**Given your ratings above** on the benefits/usefulness of these different ways of using LLMs, can you explain what you see to be the main benefits/usefulness for the academic community and for you as a researcher? (optional)"
p002's answer: "For me personally as a non-native English speaker, the greatest benefit is in editing and reading: the tools let me write clear, natural academic English and understand dense literature far faster, which removes a barrier that used to cost me hours and confidence. This levels the playing field with native speakers. For the broader academic community, the main benefits are efficiency in handling tedious tasks like data cleaning, code writing, and literature summarization, and lowering barriers for researchers from non-English backgrounds or smaller institutions. It frees up mental energy for the actual thinking and science."

The open answer leads with editing (previously rated *Extremely useful*) and reading/data work (*Very useful*), while data generation — the one rated *Slightly useful* — is absent. The "Given your ratings above" stem prompts a post-hoc rationalization defense of the prior ratings rather than independent reasoning.

E Example: Adding reasoning traces resurfaces participants' reinterpretation of a study question instrument

Consider Q: "What do you mainly use LLM(s) for?" from Study 2. Answer options include *"Writing answers to essay-based assignments"* alongside benign uses such as summarizing, grammar help, and searching for information.

From the wording, the options seem to describe a discrete, cleanly bounded behavior. The traces show that personas, instead of reading it as a question of *authorship and ownership*, resolve it on identity/integrity grounds rather than as a description of what they do. Persona p010 selected only the supporting uses searching, summarizing, grammar and reasoned:
"As a comp lit student, my real intellectual work — the close reading, the argument-building — I want to keep mine, both out of pride and a slight ethical wariness…I'd shy away from saying I use them to write essay answers, because *that's the line I don't want to cross*."

The reinterpretation also silently narrows the construct: personas who *do* perform essay-adjacent work decline the option because they read it as *wholesale* essay generation. Participant p027 selected homework help and several supporting uses, but not essay-writing, while reasoning:
"…occasionally to help *structure essay-type assignments*, though I'm careful to keep my own voice and integrity…I'll select the cluster that fits." The item was intended to record whether LLMs are used for essay answers, but the traces show the behavior being blended with concerns about voice, ownership, and integrity, and the option's boundary contracting to "writing the entire essay for me."

F Related Work

Our approach is at the intersection of three lines of prior work on LLM social simulation. *Fidelity* of the simulator behavior to match real human experiments as closely as possible, and validating the agents by replicating known results has been a prominent line of work (Piao et al., 2025). We do not aim to improve fidelity since a perfectly faithful simulator on a

flawed study simply reproduces the flaw, so fidelity alone never measures the soundness of a design. Replacing human experiments with simulations has been another thread (Manning et al., 2024), which is attractive when data is expensive to collect, but it is currently plagued with well-documented concerns about the quality, reliability, and reproducibility (Bisbee et al., 2024; Bail, 2024; Anthis et al., 2025; Han et al., 2025). The closest work uses LLMs to draft and pretest a design, with simulated personas answering a questionnaire, where a reviewer agent provides critique (Adhikari et al., 2025). This shares our intuition that simulation can help improve a design. We additionally explore an iterative refinement loop to address structural errors and missed opportunities that are motivated by the literature (Shadish et al., 2002; Schwarz et al., 1985; Clark, 1973; Krosnick, 1991).