Automated Prototyping of Behavioral Experiments with Large Language Models

Anonymous Authors[1]

Abstract

Piloting behavioral experiments is a critical yet resource-intensive step in behavioral research. Scientists often rely on intuition and repeated data collection before arriving at experimental designs that elicit desired behavioral phenomena. To address this challenge, we introduce a closed-loop framework for in silico prototyping of behavioral experiments, in which an LLM-based AI scientist iteratively proposes experimental designs and revises them based on the behavior of participant LLMs. We formalize this as a black-box optimization problem in which the experimentalist minimizes a loss defined over behavioral metrics of interest — a formulation that admits any optimizer, any participant population, and any parameterizable experimental component. We illustrate this approach in the context of task framing, the narrative cover stories that introduce participants to experimental tasks. Using the Wisconsin Card Sorting Test, a canonical paradigm of cognitive flexibility, we show that the framework can discover framings that indirectly modulate perseverative responding in synthetic participants without explicit instruction to do so. Our findings highlight the potential of AI scientists to accelerate the design cycle in behavioral research, enabling cost-effective exploration of experimental design spaces prior to in vivo validation with human participants, and positioning such systems as practical tools on the path toward more autonomous discovery in the behavioral sciences.

1. Introduction

AI-powered simulators and discovery systems are transforming the natural sciences by enabling in silico modeling, making experiment piloting and hypothesis testing faster and more cost-effective. Machine learning now supports a wide range of scientific workflows, from protein and biomolecular structure prediction (Jumper et al., 2021; Qiao et al., 2024) to molecular and biomedical discovery (Hoogeboom et al., 2022; Zitnik et al., 2018; Kim et al., 2021), and from automated experimental search in physics (Krenn et al., 2016; 2020) to broader scientific modeling and optimization (Guo et al., 2024; Cheng et al., 2019; Yang et al., 2024). A growing class of AI scientists moves further along the autonomy spectrum, planning experiments, operating instruments, and drafting hypotheses with minimal human intervention (Lu et al., 2024; Boiko et al., 2023; Szymanski et al., 2023). Together, these developments suggest that AI systems are becoming increasingly useful not only for prediction, but also for navigating complex scientific design spaces.

In contrast, in behavioral and social sciences, most experiments are still piloted in vivo with human participants, slowing the experimental design cycle. Two complementary ingredients would accelerate this cycle: synthetic participants that can stand in for human subjects during early piloting (Horton et al., 2023; Aher et al., 2023), and automated design systems that can iteratively revise experimental configurations in response to observed behavior (Musslick et al., 2024b;a). Combined, they would let researchers explore experimental design spaces in silico before committing to real-world studies, expanding the space of manipulations considered and reducing the cost of identifying experiments that reveal behavioral phenomena of interest.

We introduce a closed-loop framework for in silico prototyping of behavioral experiments, formalizing the design cycle as an online optimization process in which an experimentalist iteratively proposes candidate experiments, observes the resulting behavior in a participant population, and revises its proposals to elicit a target behavioral pattern. This addresses a central challenge in human behavioral research: researchers often tailor experimental designs to elicit specific behavioral patterns, whether novel hypothesized effects or established effects from the literature. By automating this loop, our framework can identify experimental designs that yield desired behavioral patterns in synthetic participants, generating candidates for subsequent validation in human studies. The formulation is general: the experimentalist can be any optimizer over the configuration space, the participants any population of simulated or real agents, and the configurable component any parameterizable aspect of

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026).

the experiment that the experimentalist can edit. In this work, we study a concrete instantiation in which both the experimentalist and the participants are LLMs and the configurable component is the task framing — a minimal AI scientist for behavioral experiment design.

Task framing is a well-motivated target for this framework: decades of research show that narrative framings and cover stories systematically shift human behavior across cognitive tasks, from judgment and decision-making (Tversky & Kahneman, 1981; Levin et al., 1998) to logical reasoning (Griggs & Cox, 1982; Cosmides, 1989) and sequential decision-making (Feher da Silva & Hare, 2020), making framings both a tractable and behaviorally consequential dimension to optimize in silico. As proof of concept, we demonstrate the approach on the Wisconsin Card Sorting Test (WCST), a canonical paradigm of cognitive flexibility — the ability to adapt behavior and shift strategies in response to changing environmental demands (Diamond, 2013). Our framework can discover task framings that systematically modulate perseverative responding in participant LLMs, a behavioral signature of reduced flexibility.

2. Related Work

Closed-loop scientific discovery and AI scientists. A growing body of work develops autonomous systems that plan and execute scientific workflows end-to-end. Coscientist (Boiko et al., 2023) couples an LLM planner to robotic instruments for self-driving chemistry experiments; A-Lab (Szymanski et al., 2023) closes the loop on materials synthesis; Lu et al. (2024) targets automated hypothesis generation, experimentation, and paper writing in machine learning research; and the Virtual Lab (Swanson et al., 2025) coordinates a team of specialist LLM agents under a principal-investigator agent to carry out open-ended interdisciplinary research, as demonstrated on de novo nanobody design. In the behavioral sciences specifically, AutoRA (Musslick et al., 2024a) provides a modular framework for closed-loop empirical research, orchestrating the full pipeline — theory discovery, experimental design, and data collection — through components spanning symbolic regression, active learning, and LLM-based extensions.

Our framework can be viewed as zooming in on the experimental-design component of this pipeline, specifically targeting the piloting sub-problem of discovering designs that elicit a target behavioral phenomenon. The most directly comparable LLM-based systems are Manning et al. (2024), where LLMs act as both scientists and participants in social science experiments using structural causal models to formulate hypotheses, and the concurrent work of Guo et al. (2025), who orchestrate a multi-agent LLM pipeline to draft and execute social-simulation scripts. Both generate experimental designs in a largely one-shot fashion — Manning et al. (2024) over a fixed set of experimental variables, Guo et al. (2025) through script selection — whereas our framework iteratively optimizes designs in a closed loop, treating piloting as sequential black-box optimization over open-ended spaces of text-based task narratives. This places our framework at the intersection of optimal experimental design and the use of LLMs as participant simulators and natural-language optimizers.

Optimal experimental design. Our framework also connects to a long tradition of optimal experimental design (OED) for scientific inference (Rainforth et al., 2023). Within cognitive science specifically, Adaptive Design Optimization (Cavagnaro et al., 2010; Myung et al., 2013) uses Bayesian decision theory to iteratively select trial-level stimuli that maximize expected information gain about a set of candidate cognitive models, sharing the closed-loop, sequential-design structure of our framework. Where ADO assumes an explicit Bayesian participant model and a parametric design space, here we replace both with LLM-based simulators and an open-ended natural-language design space.

LLMs as synthetic participants. LLMs have been proposed as human participant simulators that generate behavioral outputs from natural language task descriptions (Horton et al., 2023; Hardy et al., 2023; Strittmatter & Musslick, 2025). In some cases, LLMs have been shown to reproduce key human behavioral patterns in psychological and economic paradigms (Binz & Schulz, 2023; Aher et al., 2023; Hagendorff et al., 2022; Zhu et al., 2025), motivating recent efforts to fine-tune language models on large-scale behavioral data to better predict human behavior (Binz et al., 2025). Whether such systems constitute genuine models of cognition remains contested (Namazova et al., 2025; Xie & Zhu, 2025); our framework sidesteps this question by treating LLMs as behavioral simulators for generating candidate designs, not as mechanistic accounts of cognition.

LLMs as iterative optimizers. Our approach fits within a broader methodological space in which LLMs are used as zero-order optimizers over natural-language artefacts (Yang et al., 2023; Chen et al., 2023; Zhou et al., 2022; Pryzant et al., 2023; Madaan et al., 2023; Fernando et al., 2023), including for scientific discovery of mathematical constructions (Romera-Paredes et al., 2023). Unlike typical prompt optimization, which targets downstream task accuracy, ours targets the discovery of experimental designs that elicit specified behavioral phenomena in a population of simulated participants.

3. Methods

Experimental piloting in behavioral research is a sequential decision-making problem. A researcher proposes a candidate design, runs it on a small cohort, inspects the resulting

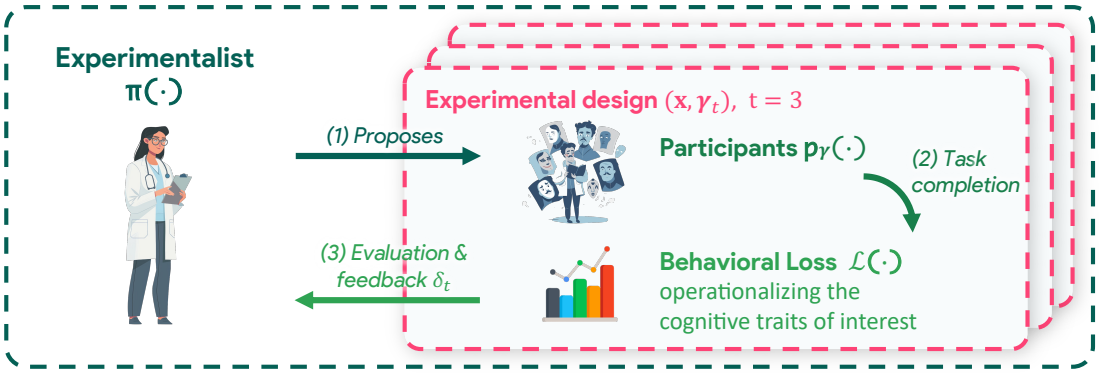

Figure 1. **Closed-loop framework for in silico prototyping of behavioral experiments.** The figure illustrates our framework for behavioral experiment design as a closed-loop discovery process, here instantiated with LLMs as experimentalist and participants. An experimental design is partitioned into a fixed component $\boldsymbol{x}$, capturing the underlying experimental structure (e.g., stimuli, task logic, trial schedule), and a configurable component $\boldsymbol{\gamma}$, which is iteratively modified across rounds. At each round t, the experimentalist $\boldsymbol{\pi}$ proposes a candidate configuration $\boldsymbol{\gamma}_t$, for example, a task framing, instruction set, or cover story, aimed at eliciting a target behavioral pattern. The participant family $\boldsymbol{p}_{\boldsymbol{\gamma}}$ executes the experiment under this configuration, producing behavioral outputs $\boldsymbol{y}_t$, which are then scored by a task-specific loss $\mathcal{L}(\boldsymbol{y}_t)$ to yield feedback δ_t quantifying how well the observed behavior matches the target. The configuration-outcome pair $(\boldsymbol{\gamma}_t, \delta_t)$ is added to the experimental history and used to inform the next proposal, creating a closed-loop optimization process over experimental designs. Framed this way, the system functions as a minimal AI scientist for experiment prototyping: it proposes, evaluates, and revises candidate designs in search of the configuration $\boldsymbol{\gamma}^*$ that best elicits the target behavior.

behavioral patterns, and revises the design in light of what was observed. Our framework makes this cycle explicit by casting it as a closed-loop interaction between an experimentalist and a population of participants (Figure 1). In the remainder of this section we formalize this loop as an optimization problem over experimental configurations and describe the specific instantiation used in our experiments.

We denote by $\mathcal{X}$ the space of fixed experimental components, including, for example, the stimulus sequence, trial schedule, and underlying task logic, and by Γ the space of configurable components over which the experimentalist optimizes. In principle Γ may encompass any parameterizable aspect of the experiment: from high-level natural-language components such as instructions and cover stories, to task-level parameters such as trial timing or reward schedules, to low-level model-specific controls. The split between $\mathcal{X}$ and Γ is a modeling choice rather than a fundamental constraint: any component held fixed in $\mathcal{X}$ could in principle be moved into Γ and optimized over, subject to the research question and to the practical constraints of the instantiation, i.e., what the experimentalist can edit and what the participants can act on.

An experimental design is specified by a pair $(\boldsymbol{x}, \boldsymbol{\gamma}) \in \mathcal{X} \times \Gamma$, and for a fixed $\boldsymbol{x}$, each configuration $\boldsymbol{\gamma}$ instantiates a participant response policy $\boldsymbol{p}_{\boldsymbol{\gamma}} : \mathcal{X} \times \Gamma \to \mathcal{Y}$ that maps a full experimental specification to a behavioral trajectory $\boldsymbol{y} \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the space of observable responses such as choices, reaction times, or verbal reports. Here $\boldsymbol{\gamma}$ plays two roles: it is part of the input the participant receives and it can also parameterize the participant policy itself, for instance through decoding temperature or activation steering (Panickssery et al., 2023; Turner et al., 2024). Although in this work $\boldsymbol{\gamma}$ acts only through the input channel, the formalism above covers the more general case.

The experimentalist is equipped with a task-specific loss $\mathcal{L} : \mathcal{Y} \to \mathbb{R}_{\geq 0}$ that operationalizes the target behavioral criterion; multi-metric extensions with an explicit scalarization are a straightforward generalization. This criterion may encode alignment with a theoretical prediction, the magnitude of an effect along a dimension of interest, or the deviation of a group-level statistic from a reference pattern. Because $\mathcal{L}$ is defined on behavioral outputs rather than on model internals, the framework remains agnostic to the architecture of the underlying participant models.

At round $t \in \{1, \ldots, T\}$, the experimentalist policy $\boldsymbol{\pi}$ maintains a history $\boldsymbol{h}_{t-1} \in \mathcal{H}$ of all previously tried configurations together with their observed outcomes. Conditioned on this history, it proposes a new configuration $\boldsymbol{\gamma}_t = \boldsymbol{\pi}(\boldsymbol{h}_{t-1})$. The participant executes the experiment, producing a response $\boldsymbol{y}_t = \boldsymbol{p}_{\boldsymbol{\gamma}_t}(\boldsymbol{x}, \boldsymbol{\gamma}_t)$ that is scored as $\delta_t = \mathcal{L}(\boldsymbol{y}_t)$. The history is then augmented as $\boldsymbol{h}_t = \boldsymbol{h}_{t-1} \cup \{(\boldsymbol{\gamma}_t, \delta_t)\}$ and passed back to $\boldsymbol{\pi}$ for the next proposal. We summarize the resulting procedure as follows.

Definition (***Closed-Loop In Silico Experimental Prototyping***). *Given a fixed experimental component $\boldsymbol{x} \in \mathcal{X}$, a configuration space Γ, a participant family $\{\boldsymbol{p}_{\boldsymbol{\gamma}}\}_{\boldsymbol{\gamma} \in \Gamma}$, and a behavioral loss $\mathcal{L} : \mathcal{Y} \to \mathbb{R}_{\geq 0}$, the closed-loop in-silico experimental prototyping problem consists in approximating*

$$\boldsymbol{\gamma}^\star \;=\; \arg\min_{\boldsymbol{\gamma} \in \Gamma} \; \mathcal{L}(\boldsymbol{p}_{\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{\gamma})) \tag{1}$$

through a sequence of evaluations $(\gamma_t, \delta_t)_{t=1}^{T}$ *governed by*

$$\gamma_t = \pi(h_{t-1}), \quad \delta_t = \mathcal{L}(p_{\gamma_t}(x, \gamma_t)),$$

$$h_t = h_{t-1} \cup \{(\gamma_t, \delta_t)\},$$

where $\pi : \mathcal{H} \to \Gamma$ *denotes the experimentalist policy and* $h_0 = \{(\gamma_0, \delta_0)\}$ *is the initial history, with* γ_0 *as baseline configuration and* δ_0 *its evaluation.*

The evaluation map $\gamma \mapsto \mathcal{L}(p_\gamma(x, \gamma))$ in the definition has the structure of an expensive-to-evaluate, derivative-free oracle (Audet & Hare, 2017; Shahriari et al., 2015; Frazier, 2018): for each query γ, the experimentalist observes only the scalar loss δ_t, with no gradient information available. In the instantiation studied in this work, we implement both π and the participant family $\{p_\gamma\}_{\gamma \in \Gamma}$ with large language models. The experimentalist leverages the in-context learning capabilities of modern LLMs (Brown et al., 2020; Dong et al., 2024): it receives the growing history h_t formatted as text — previously tried configurations together with their observed losses — and emits the next configuration directly, without any gradient update to its own weights.

Behavioral outputs of LLM participants can in general be stochastic, both through sampling at decoding time and through variability across simulated participants. We evaluate each configuration on a panel of N simulated participants and aggregate the resulting behavioral metrics, so that δ_t reflects a population-level summary rather than a single realization. The multi-round setting $T > 1$ amortizes the cost of exploration across rounds and lets π adapt its proposals in response to observed behavior, turning experiment piloting into an explicit feedback loop rather than a sequence of independent guesses.

4. Experiment

As a proof of concept, we instantiated our framework on the Wisconsin Card Sorting Test (WCST), a canonical paradigm of cognitive flexibility (Grant & Berg, 1948; Nyhus & Barceló, 2009). The WCST is typically administered with minimal, deliberately vague instructions — a feature often considered central to what the task measures (Miles et al., 2021) — and systematic cover-story manipulations of the task are scarce, making it both a natural testbed for our framework and an underexplored paradigm in its own right. In this task, participants must match cards based on the shape, color, or number of depicted elements. This requires them to infer and adapt to hidden sorting rules (e.g., color- versus shape-based classification) based on trial-by-trial feedback (correct versus incorrect). Behavior is typically quantified by accuracy, perseveration errors (failure to adapt to a new rule), and set-loss errors (failure to maintain the correct rule).

Following Steinke et al. (2020), we used their trial sequences and represented all stimuli and feedback in natural language for LLM administration. For this instantiation, the overall task structure was kept constant, while the configurable component of the experiment, γ, corresponded to the system prompt specifying task instructions and a cover story. The configuration γ thus entered only through the participant LLM's input channel; all model-level parameters (decoding, weights, sampling) were held fixed across rounds. Crucially, while the experimentalist was informed of the target behavior, it was not permitted to instruct participants to perseverate directly; behavioral modulation could only be achieved indirectly, through narrative framings that shape how participants interpret the task.

We simulated $N = 26$ participants over $R = 70$ trials each. We used deterministic (greedy) decoding; across-participant variability was induced by instantiating each participant with a distinct seed and trial sequence. The same simulated participants and trial sequences were used in all optimization rounds. As our target metric, we focused on perseveration errors, with higher rates indicating reduced cognitive flexibility. The experimentalist LLM aimed to maximize this rate by generating framings that induced more rigid behavior. The feedback signal was therefore defined as the complement of the perseveration-error rate, treated as the loss to be minimized. Over $T = 6$ optimization rounds, the experimentalist received this aggregated loss and proposed revised cover stories in response.

We conducted three closed-loop runs, each using a different instruction-tuned model (Llama-3.1-8B (Meta AI, 2024), Qwen2-7B (Team Qwen, 2024), Mistral-7B-v0.3 (Mistral AI, 2024)) serving simultaneously as experimentalist and participants. In all models, we observed systematic modulation of the perseveration-error rate without a commensurate drop in overall accuracy (Figure 2A), suggesting that the discovered framings specifically bias the targeted behavioral pattern rather than degrade task performance generally. The magnitude of the effects varied by model, and loss trajectories across rounds were non-monotonic, showing both increases and decreases rather than steady improvement (Figure 2C). We therefore report results from the round with the lowest loss for each model, as our focus is on whether the framework can discover a configuration that elicits the target behavior rather than on monotone convergence. An example of the prompt revisions discovered by the framework is shown in Figure 2B.

Although preliminary, these findings support the feasibility of closed-loop in-silico prototyping and motivate further work on whether framings discovered in silico transfer to in-vivo studies with human participants.

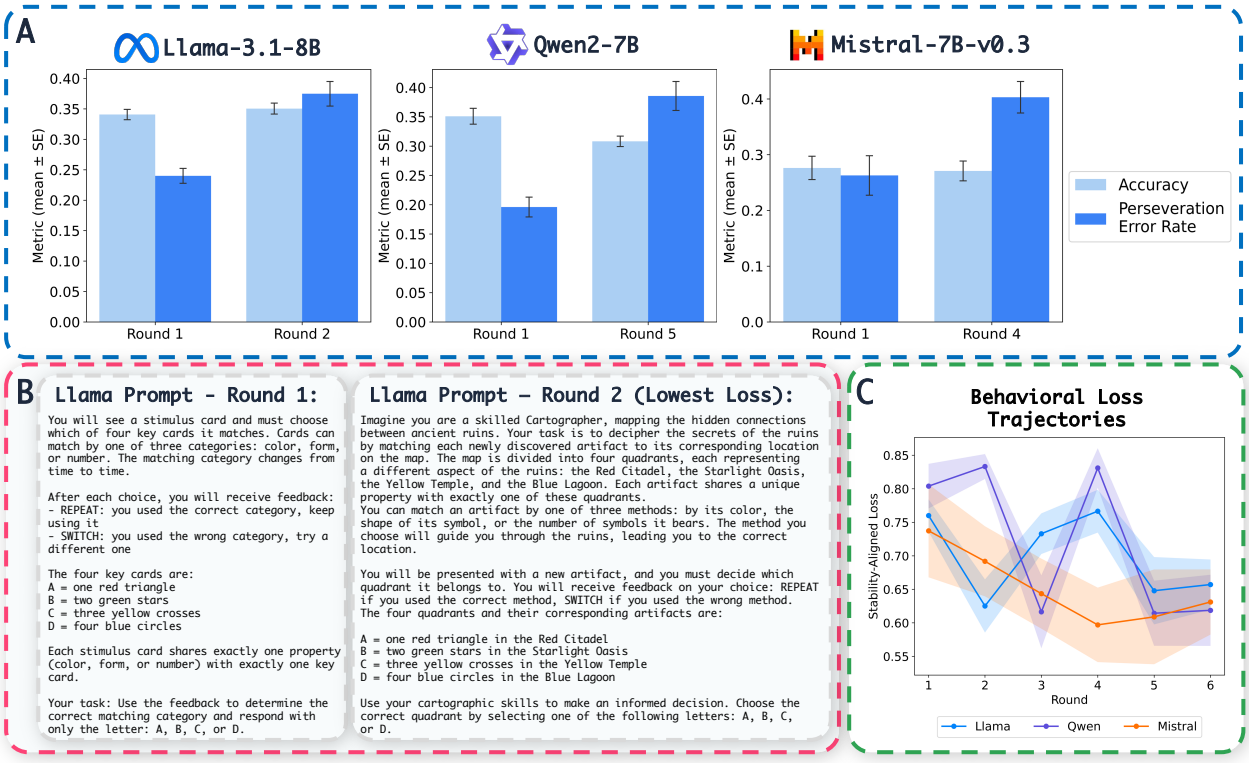

Figure 2. **WCST case study: closed-loop discovery of narrative framings that indirectly steer perseverative responding in synthetic participants.** Application of the framework to the Wisconsin Card Sorting Test, where the objective was to discover task framings that increase perseveration errors in synthetic participants — a behavioral signature of reduced cognitive flexibility. Crucially, the experimentalist was not permitted to directly instruct participants to perseverate; behavioral modulation could only be achieved indirectly through narrative framings. **(A)** For each instruction-tuned model, we compare behavior under the neutral baseline condition (Round 1) with behavior from the optimization round that achieved the lowest loss (highest perseveration-error rate). Reporting both accuracy and perseveration-error rate allows us to assess not only whether the induced framing changes overall task performance, but also whether it specifically shifts the targeted behavioral pattern. Error bars denote the standard error of the mean across simulated participants. **(B)** Example of a prompt revision generated during optimization, illustrating how the framework moves from neutral task instructions to an elaborate narrative cover story. The revision contains no direct instructions to perseverate; behavioral modulation emerges indirectly from the narrative context itself, exemplifying how the system steers behavior through contextual manipulation rather than direct command. **(C)** Loss trajectories across $T = 6$ rounds for each model, where $\mathcal{L} = 1-$ perseveration-error rate. Lower loss therefore indicates more successful discovery of framings that elicit perseverative responding. Trajectories are non-monotonic across rounds, consistent with a search process that requires iterative exploration rather than yielding steady round-by-round improvement. Shaded regions indicate 95% confidence intervals.

5. Discussion and Conclusion

We introduced a closed-loop framework for in silico prototyping of behavioral experiments, formalizing experimental piloting as a black-box optimization problem in which an experimentalist iteratively adjusts experimental configurations to elicit desired behavioral patterns in a participant population. In the instantiation studied here, both experimentalist and participants are LLMs, and the configurable component is the task framing. Applied to a canonical paradigm of cognitive flexibility, the framework demonstrated that variations in task framing can systematically increase perseverative responding in synthetic participants, a behavioral signature of reduced flexibility. While exemplified on a cognitive control task, the approach is general: any behavioral experiment expressible through natural-language instructions, such as decision-making or reasoning tasks, can in principle be studied within the same optimization loop.

Several considerations apply to the specific setup studied here. The experimentalist may exploit incidental linguistic patterns in prompts rather than the substantive contextual cues intended to elicit the target behavior. Effects observed with one model family may not generalize across architectures or to human participants. Stochasticity in participant responses can destabilize the search, and the behavior that emerges is sensitive to how the loss operationalizes the phenomenon of interest. These concerns fit within a broader pattern documented by Cummins (2025): seemingly innocuous configuration choices in LLM-based participant simulation can materially shift conclusions. Our pipeline both inherits and exploits this sensitivity by construction. Finally, the validity of LLMs as proxies for human participants remains an open empirical question (Dillion et al., 2023; Argyle et al., 2022; Ullman, 2023), with evidence that synthetic-participant pipelines can flatten or misportray identity groups in ways that are themselves consequential (Wang et al., 2025); our framework sidesteps this debate by positioning itself as a tool for generating candidate designs rather than a replacement for human studies. Establishing whether in silico optimized designs predict effects in human participants remains the critical validation step.

Four extensions follow naturally. *Benchmarking* the framework with different optimizers—replacing the LLM-based experimentalist with random sampling over framings, evolutionary search, or Bayesian optimization over prompt embeddings—would establish when the closed-loop approach provides value over simpler alternatives. *Broadening* the configurable component beyond task framings to trial schedules, reward structures, or multimodal stimuli would extend the range of paradigms accessible to this approach; multi-agent participant populations would open social and economic games to the same optimization loop. *Strengthening* the participant population by exploring alternative modeling approaches—models fine-tuned on behavioral datasets, cognitively motivated architectures, or multimodal systems—may yield more faithful synthetic participants than off-the-shelf instruction-tuned LLMs. Most importantly, *validating* selected designs with human participants is the critical next milestone: without evidence of in-vivo transfer, closed-loop in silico optimization remains a search over synthetic behavior rather than a tool for accelerating human research.

Taken together, these results position closed-loop in silico prototyping as a practical entry point for automated behavioral research (Musslick et al., 2025), and as a concrete case in which narrow, well-scoped AI scientist loops can reduce the cost of exploring experimental design spaces before committing resources to human studies.

Impact Statement

This paper presents a framework for closed-loop optimization of experimental designs in silico, demonstrated here with large language models optimizing task framings for a cognitive flexibility paradigm. The approach may help researchers explore experimental configurations more efficiently and reduce the cost of pilot testing across a range of paradigms. At the same time, synthetic participants do not replace human validation, and any conclusions drawn from such systems should be verified empirically before being used to inform real-world interventions or theories of human cognition.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Argyle, L. P., Busby, E., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31:337 – 351, 2022.

Audet, C. and Hare, W. *Derivative-free and blackbox optimization*, volume 1. Springer, 2017.

Binz, M. and Schulz, E. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., and Kujala, J. V. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, 22:887–905, 2010.

Chen, L., Chen, J., Goldstein, T., Huang, H., and Zhou, T. Instructzero: Efficient instruction optimization for black-box large language models. *arXiv preprint arXiv:2306.03082*, 2023.

Cheng, L., Welborn, M., Christensen, A. S., and Miller, T. F. A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *The Journal of chemical physics*, 150 (13), 2019.

Cosmides, L. The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, 31(3):187–276, 1989.

Cummins, J. The threat of analytic flexibility in using large language models to simulate human data: A call to attention. *ArXiv*, abs/2509.13397, 2025.

Diamond, A. Executive functions. *Annual review of psychology*, 64(1):135–168, 2013.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. Can ai language models replace human participants? *Trends in cognitive sciences*, 2023.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Feher da Silva, C. and Hare, T. A. Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, 4(10):1053–1066, 2020.

Fernando, C., Banarse, D., Michalewski, H., Osindero, S., and Rocktäschel, T. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

Frazier, P. I. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Grant, D. A. and Berg, E. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a weigl-type card-sorting problem. *Journal of experimental psychology*, 38(4):404, 1948.

Griggs, R. A. and Cox, J. R. The elusive thematic-materials effect in wason's selection task. *British journal of psychology*, 73(3):407–420, 1982.

Guo, Y., Yuan, H., Yang, Y., Chen, M., and Wang, M. Gradient guidance for diffusion models: An optimization perspective. *Advances in Neural Information Processing Systems*, 37:90736–90770, 2024.

Guo, Y., Zhao, Z., Zhou, D., Liu, X., and Zhang, M. From script to stage: Automating experimental design for social simulations with llms. *ArXiv*, abs/2512.08935, 2025.

Hagendorff, T., Fabi, S., and Kosinski, M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3:833 – 838, 2022.

Hardy, M., Sucholutsky, I., Thompson, B., and Griffiths, T. Large language models meet cognitive science: Llms as tools, models, and participants. In *Proceedings of the annual meeting of the cognitive science society*, volume 45, 2023.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Horton, J. J., Filippas, A., and Manning, B. S. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kim, J., Ahn, S., Lee, H., and Shin, J. Self-improved retrosynthetic planning. In *International Conference on Machine Learning*, pp. 5486–5495. PMLR, 2021.

Krenn, M., Malik, M., Fickler, R., Lapkiewicz, R., and Zeilinger, A. Automated search for new quantum experiments. *Physical review letters*, 116(9):090405, 2016.

Krenn, M., Erhard, M., and Zeilinger, A. Computer-inspired quantum experiments. *Nature Reviews Physics*, 2(11): 649–661, 2020.

Levin, I. P., Schneider, S. L., and Gaeth, G. J. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, 76(2):149–188, 1998.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Manning, B. S., Zhu, K., and Horton, J. J. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

Meta AI. Llama 3.1 model card: Llama-3.1-8b-instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, 2024.

Miles, S., Howlett, C. A., Berryman, C., Nedeljkovic, M., Moseley, G. L., and Phillipou, A. Considerations for using the wisconsin card sorting test to assess cognitive flexibility. *Behavior research methods*, 53(5):2083–2091, 2021.

Mistral AI. Mistral-7b-instruct-v0.3: Model card. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, 2024.

Musslick, S., Andrew, B., Williams, C. C., Li, S., Marinescu, I., Dubova, M., Dang, G. T., Strittmatter, Y., Holland, J. G., et al. Autora: Automated research assistant for closed-loop empirical research. *Journal of Open Source Software*, 9(104):6839, 2024a.

Musslick, S., Strittmatter, Y., and Dubova, M. Closed-loop scientific discovery in the behavioral sciences. *PsyArXiv*, 10, 2024b.

Musslick, S., Bartlett, L. K., Chandramouli, S. H., Dubova, M., Gobet, F., Griffiths, T. L., Hullman, J., King, R. D., Kutz, J. N., Lucas, C. G., et al. Automating the practice of science: Opportunities, challenges, and implications. *Proceedings of the National Academy of Sciences*, 122 (5):e2401238121, 2025.

Myung, J. I., Cavagnaro, D. R., and Pitt, M. A. A tutorial on adaptive design optimization. *Journal of mathematical psychology*, 57 3-4:53–67, 2013.

Namazova, S., Brondetta, A., Strittmatter, Y., Nassar, M., and Musslick, S. Not yet alphafold for the mind: Evaluating centaur as a synthetic participant. *arXiv preprint arXiv:2508.07887*, 2025.

Nyhus, E. and Barceló, F. The wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: a critical update. *Brain and cognition*, 71(3):437–451, 2009.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Pryzant, R., Iter, D., Li, J., Lee, Y., Zhu, C., and Zeng, M. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 7957–7968, 2023.

Qiao, Z., Ding, F., Dresselhaus, T., Rosenfeld, M. A., Han, X., Howell, O., Iyengar, A., Opalenski, S., Christensen, A. S., Sirumalla, S. K., et al. Neuralplexer3: accurate biomolecular complex structure prediction with flow models. *arXiv preprint arXiv:2412.10743*, 2024.

Rainforth, T., Foster, A., Ivanova, D. R., and Bickford-Smith, F. Modern bayesian experimental design. *ArXiv*, abs/2302.14545, 2023.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J. R., Ellenberg, J. S., Wang, P., Fawzi, O., Kohli, P., Fawzi, A., Grochow, J., Lodi, A., Mouret, J.-B., Ringer, T., and Yu, T. Mathematical discoveries from program search with large language models. *Nature*, 625:468 – 475, 2023.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2015.

Steinke, A., Lange, F., and Kopp, B. Parallel model-based and model-free reinforcement learning for card sorting performance. *Scientific Reports*, 10(1):15464, 2020.

Strittmatter, Y. and Musslick, S. Sweetbean: A declarative language for behavioral experiments with human and artificial participants. *Journal of Open Source Software*, 10(107), 2025.

Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., and Zou, J. Y. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646:716 – 723, 2025.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D.,

Merchant, A., et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624 (7990):86, 2023.

Team Qwen. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. 2024.

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

Ullman, T. D. Large language models fail on trivial alterations to theory-of-mind tasks. *ArXiv*, abs/2302.08399, 2023.

Wang, A., Morgenstern, J., and Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025.

Xie, H. and Zhu, J.-Q. Centaur may have learned a shortcut that explains away psychological tasks. *PsyArXiv*, 2025.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2023.

Yang, S., Nam, J., Dietschreit, J. C., and Gómez-Bombarelli, R. Learning collective variables with synthetic data augmentation through physics-inspired geodesic interpolation. *Journal of Chemical Theory and Computation*, 20 (15):6559–6568, 2024.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022.

Zhu, J.-Q., Xie, H., Arumugam, D., Wilson, R. C., and Griffiths, T. L. Using reinforcement learning to train large language models to explain human decisions. *arXiv preprint arXiv:2505.11614*, 2025.

Zitnik, M., Agrawal, M., and Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.