# Automated Prototyping of Behavioral Experiments with Large Language Models

**Alessandra Brondetta** *Osnabrück University* albrondetta@uni-osnabrueck.de
**Sebastian Musslick** *Osnabrück University, Brown University*
sebastian.musslick@uos.de

## Abstract

Piloting behavioral experiments is a critical yet resource-intensive step in behavioral research. Behavioral 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 large language model (LLM)-driven framework for in silico prototyping of behavioral experiments. The framework involves an iterative interaction between an experimentalist LLM, that proposes candidate designs, and participant LLMs, that engage with them. We formalize this interaction as a black-box optimization problem, where the experimentalist LLM aims to minimize a loss function defined over behavioral metrics of interest by iteratively revising its proposals. We illustrate this approach in the context of task framing—the narrative explanations used to introduce participants to experimental tasks. Using the Wisconsin Card Sorting Test, a canonical psychological paradigm for studying cognitive flexibility, we demonstrate that the framework can discover framings that systematically shift the behavior of synthetic participants along a spectrum of cognitive stability and flexibility. Our findings demonstrate the potential of LLM-based in silico experimentation to accelerate the design cycle in behavioral research, enabling cost-effective exploration of experimental design spaces prior to in vivo validation with human participants.

**Keywords:** cognitive science, automated scientific discovery, optimal experimental design

## 1 Introduction

AI-powered simulators have transformed the natural sciences by enabling in silico modeling, making experiment piloting and hypothesis testing faster and more cost-effective. For example, high-accuracy predictions of protein structure as accomplished by AlphaFold accelerated experiment prototyping and hypothesis testing in chemistry [Jumper et al., 2021]. In contrast, in behavioral and social sciences, most experiments are still piloted in vivo with human participants, slowing the experimental design cycle. A "synthetic participant" would let researchers explore experimental designs in silico before committing to real-world studies [Musslick et al., 2024], thus accelerating iteration, expanding design spaces, and reducing the costs associated with identifying experiments that reveal behavioral phenomena of interest.

We introduce a framework for in silico prototyping of behavioral experiments using large language models (LLMs) as both synthetic participants and exper-

iment designers. The framework formalizes their interaction as a closed-loop, online, optimization process, enabling automated exploration of experimental conditions that shape behavior. This addresses a central challenge in human behavioral research: researchers often tailor experimental designs to elicit specific behavioral patterns, such as novel hypothesized effects or existing effects from the literature, e.g., for testing interactions with other factors. Our framework automates this process, enabling the identification of experimental designs that yield desired behavioral patterns in synthetic participants, thereby generating candidate designs that may translate to human studies.

To make this concrete, consider a researcher investigating how acute stress influences cognitive flexibility–the ability to adapt behavior and shift strategies in response to changing environmental demands [Diamond, 2014]. Instead of conducting multiple costly pilot studies with human participants to fine-tune stress-level manipulations, our framework can systematically explore, in silico, narrative framings and task features (e.g., time pressure, competitive instructions, task difficulty) to identify which combinations elicit the desired stress responses and corresponding changes in flexibility metrics among LLM-based participants. Through iterative feedback, the experimentalist LLM explores task configurations based on observed synthetic behaviors, narrowing down candidate designs that may later be validated with human participants and accelerating the early stages of experiment design.

As proof of concept, we demonstrate the approach on the Wisconsin Card Sorting Test (WCST), a canonical paradigm of cognitive flexibility, showing how our method discovers task framings that systematically modulate the behavior of participant LLMs along their flexibility–stability spectrum.

## 2 Relevant Work

LLMs have been proposed as human participant simulators that generate behavioral outputs from natural language task descriptions [Hardy et al., 2023, Manning et al., 2024, Strittmatter and Musslick, 2025]. In some cases, LLMs have been shown to reproduce key human behavioral patterns in psychological paradigms [Binz and Schulz, 2023, Aher et al., 2023, Zhu et al., 2025]. As with other simulators in the natural sciences, an LLM need not provide a mechanistic account of cognition to be useful, but rather reliably generate observable behavior across experimental contexts [Namazova et al., 2025].

Recent work has also explored LLMs for experimental design. Manning et al. [2024] introduced a framework where LLMs act both as scientists and participants in social science experiments, using structural causal models to formulate hypotheses and design experiments. However, their approach generated experiments in a one-shot manner, restricted to fixed set of experimental variables. Our framework instead optimizes experimental design to elicit targeted behavioral phenomena, operating over open-ended spaces of text-based task narratives. This approach fits within the broader methodological space of LLMs as iterative optimizers [Yang et al., 2023, Chen et al., 2023].

# 3 Methods

Our framework for in silico experimental prototyping of behavioral experiments formalizes the interaction between an experimentalist LLM and participants LLMs as an iterative optimization process (see Figure 1). An experiment is defined by a pair $(\boldsymbol{x}, \boldsymbol{\gamma})$, where $\boldsymbol{x} \in \mathcal{X}$ denotes the fixed component of the experiment (e.g., the task structure or stimuli), and $\boldsymbol{\gamma} \in \Gamma$ denotes a configurable component (e.g., task framing, instructions or experiment duration). In each round $t$ of piloting, (1) the experimentalist $\boldsymbol{\pi}$ proposes a configuration $\boldsymbol{\gamma}_t$, (2) the participants $\boldsymbol{p}_{\boldsymbol{\gamma}}$ complete the experiment under this configuration, and (3) their responses $\boldsymbol{y}_t$ are evaluated using a task-specific loss function $\mathcal{L} : \mathcal{Y} \to \mathbb{R}_{\geq 0}$. This loss encodes a behavioral objective, for example, how closely the synthetic responses align with a target behavioral pattern or theoretical construct such as accuracy, bias, or flexibility. The resulting score $\delta_t = \mathcal{L}(\boldsymbol{y}_t)$ thus quantifies the extent to which the observed behavior matches the desired behavioral criterion. This feedback $\delta_t$, coupled with the configuration $\boldsymbol{\gamma}_t$ that produced it, is recorded in the experimental history, which influences the subsequent proposals of the experimentalist.

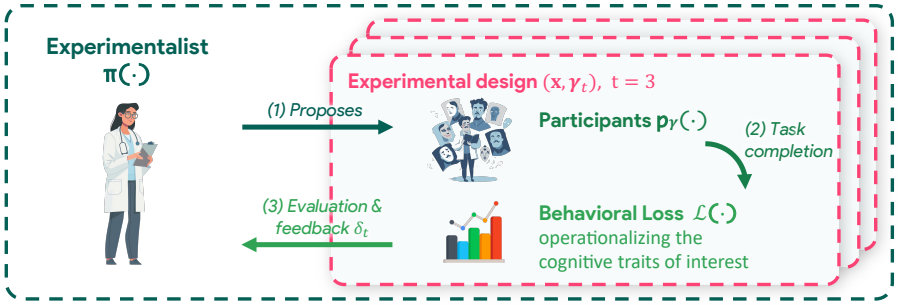


Figure 1: **In-silico experimental prototyping loop.**

For fixed $\boldsymbol{x}$, the experimentalist policy $\boldsymbol{\pi}(\cdot)$ acts as a black-box optimizer, sequentially exploring $\Gamma$ to identify the configuration $\boldsymbol{\gamma}^*$ that best elicits the target behavior, that is, the configuration yielding the minimal loss:

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma} \in \Gamma} \mathcal{L}\big(\boldsymbol{p}_{\boldsymbol{\gamma}}(\boldsymbol{x})\big).$$

# 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 and Berg, 1948, Nyhus and Barceló, 2009]. 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 card 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).

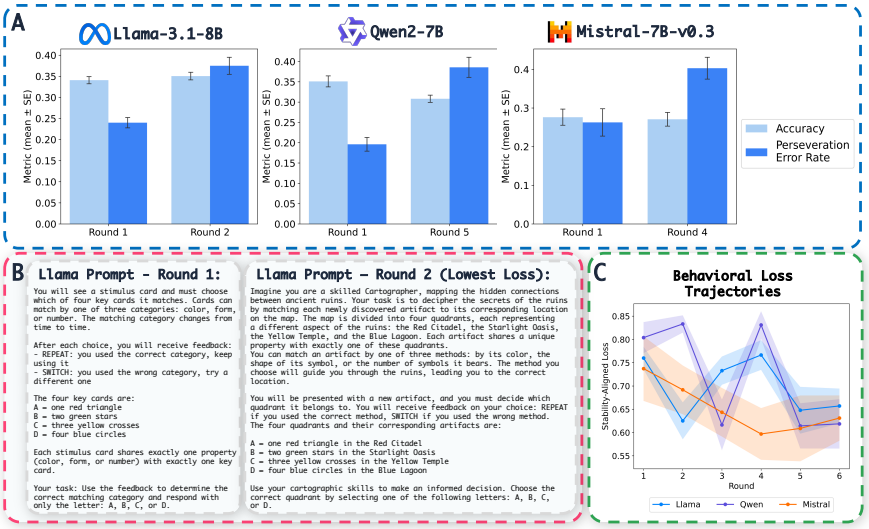


Figure 2: **WCST experimental prototyping results. (A)** For each model, accuracy and perseveration-error rate are shown for neutral baseline (Round 1) and for the round with the highest perseveration rate (lowest loss), indicating reduced flexibility. Error bars represent standard error of the mean across simulated participants. **(B)** Example prompt change from neutral instructions to a cover story illustrating contextual framing alterations between rounds for Llama. **(C)** Stability-aligned loss ($\mathcal{L} = 1-$ perseveration-error rate) across T=6 rounds for each model. Shaded areas indicate 95% confidence intervals.

We adapted the task for LLMs by representing all stimuli and feedback in natural language and using the trial sequences from Steinke et al. [2020]. For this instantiation, the overall task structure was kept constant, while the configurable component of the experiment, $\gamma_t$, corresponded to the system prompt specifying task instructions and a cover story designed to bias behavior. In this application, the optimization process therefore operated over narrative framings, although in principle other components of an experiment, such as trial length, stimuli or responses modalities, could be parameterized in the same way.

Following Steinke et al. [2020], we simulated $N = 26$ participants over $R = 70$ trials each. The same simulated participants and trial sequences were used in all optimization rounds. As our target metric, we focused on perseveration errors as an index of cognitive flexibility, with higher error rates indicating reduced flexibility. The experimentalist LLM aimed to maximize perseveration, generating framings that induced more rigid behavior among participant LLMs. The feedback signal was therefore based on the complement of the perseveration-error rate, treated as the loss to be minimized in this specific application. Over $T = 6$ optimization rounds, the experimentalist received this aggregated loss and proposed revised cover stories in response, iteratively exploring framings that modulated flexibility.

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, demonstrating that LLM-generated cover stories can influence simulated cognitive flexibility. 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. We report results from the round with the lowest loss, as our focus is on the configuration that best elicits the target behavior (Figure 2).

While preliminary, these findings validate the feasibility of closed-loop in silico prototyping and suggest the potential for in silico exploration of experimental designs prior to costly in vivo studies.

# 5 Discussion and Conclusion

We introduced an AI-driven framework for in silico prototyping of behavioral experiments, formalizing experimental piloting as a black-box optimization problem in which an experimentalist LLM iteratively adjusts experimental configurations, instantiated here as task framings, to steer participant LLMs toward desired behavioral patterns. Applied to a canonical paradigm of cognitive flexibility, the framework demonstrated that variations in task framing can systematically bias synthetic participants toward 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 tasks, or social interaction games, can in principle be instantiated within the same optimization loop. Extensions involving multi-agent setups or multimodal systems (e.g., using vision-language models for visual stimuli) may broaden applicability to tasks involving social or perceptual processes.

The framework's effectiveness, however, depends on how well LLMs capture the behavioral constructs under study. Potential limitations include overfitting to linguistic artifacts in prompts, model-specific biases that fail to generalize to human cognition, instability of the optimization process due to stochastic LLM responses, and sensitivity to the definition of the loss function used to operationalize behavioral goals. Moreover, the present study does not establish that in-silico-optimized designs directly translate to human behavior; empirical validation remains essential to assess predictive validity.

Critical next steps involve benchmarking the black-box optimization process against alternative approaches (e.g., random sampling), extending the framework's application across diverse behavioral phenomena and, crucially, as mentioned before, validating whether in silico optimized designs predict outcomes in human participants. Pending results from future analyses, this framework opens paths toward automated experimental design [Musslick et al., 2025], enabling behavioral scientists to explore complex design spaces more efficiently and at lower cost before committing resources to human studies.

## References


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

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

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

A. Diamond. Executive functions. *Handbook of clinical neurology*, 173:225–240, 2014.

D. A. Grant and E. Berg. 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.

M. Hardy, I. Sucholutsky, B. Thompson, and T. Griffiths. 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.

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

B. S. Manning, K. Zhu, and J. J. Horton. Automated social science: Language models as scientist and subjects. *arXiv preprint arXiv:2404.11794*, 2024.

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

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

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

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

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

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

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

Y. Strittmatter and S. Musslick. Sweetbean: A declarative language for behavioral experiments with human and artificial participants. *Journal of Open Source Software*, 10(107):7703, 2025.

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

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

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