# Automated Prototyping of Behavioral Experiments with Large Language Models

**Alessandra Brondetta** [1] **Sebastian Musslick** [1 2]

## 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. The discovered cover stories are optimized against synthetic participants and still require validation in humans, which we leave to future work. We position 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]Osnabrück University, Osnabrück, Germany [2]Brown University, Providence, Rhode Island, USA. Correspondence to: Alessandra Brondetta <albrondetta@uni-osnabrueck.de>.

*Accepted at ICML 2026 AI for Science Workshop.*

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., 2024). 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., 2023; 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., 2024; Zhou et al., 2022; Pryzant et al., 2023; Madaan et al., 2023; Fernando et al., 2024), including for scientific discovery of mathematical constructions (Romera-Paredes et al., 2024). 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

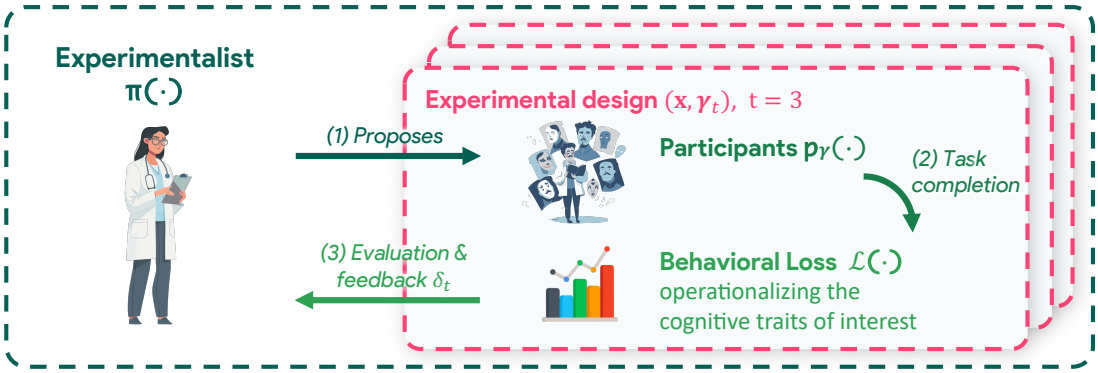


*Figure 1.* **Closed-loop framework for in silico prototyping of behavioral experiments.** The figure illustrates our framework for behavioral experiment design as a closed-loop discovery process, here instantiated with LLMs as experimentalist and participants. An experimental design is partitioned into 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 $\delta_t$ 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 $\Gamma$ the space of configurable components over which the experimentalist optimizes. In principle $\Gamma$ 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 $\Gamma$ is a modeling choice rather than a fundamental constraint: any component held fixed in $\mathcal{X}$ could in principle be moved into $\Gamma$ 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 (Rimsky et al., 2024; 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 $\Gamma$, 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* $(\boldsymbol{\gamma}_t, \delta_t)_{t=1}^{T}$ *governed by*

$$\boldsymbol{\gamma}_t = \boldsymbol{\pi}(\boldsymbol{h}_{t-1}), \quad \delta_t = \mathcal{L}(\boldsymbol{p}_{\boldsymbol{\gamma}_t}(\boldsymbol{x}, \boldsymbol{\gamma}_t)),$$

$$\boldsymbol{h}_t = \boldsymbol{h}_{t-1} \cup \{(\boldsymbol{\gamma}_t, \delta_t)\},$$

*where* $\boldsymbol{\pi} : \mathcal{H} \to \Gamma$ *denotes the experimentalist policy and* $\boldsymbol{h}_0 = \{(\boldsymbol{\gamma}_0, \delta_0)\}$ *is the initial history, with* $\boldsymbol{\gamma}_0$ *as baseline configuration and* $\delta_0$ *its evaluation.*

The evaluation map $\boldsymbol{\gamma} \mapsto \mathcal{L}(\boldsymbol{p}_{\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{\gamma}))$ 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 $\boldsymbol{\gamma}$, the experimentalist observes only the scalar loss $\delta$, without gradient information. The multi-round setting $T > 1$ amortizes exploration across rounds and lets $\boldsymbol{\pi}$ adapt its proposals to observed behavior, turning piloting into an explicit feedback loop rather than a sequence of independent guesses.

In the instantiation studied here, we implement both $\boldsymbol{\pi}$ and the participant family $\{\boldsymbol{p}_{\boldsymbol{\gamma}}\}_{\boldsymbol{\gamma} \in \Gamma}$ 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 $\boldsymbol{h}_t$ formatted as text — previously proposed configurations and their observed losses — and emits the next configuration directly, without any gradient update to its own weights. Alongside each loss, the experimentalist is also shown two read-only diagnostics, overall accuracy and invalid-response rate, which flag when a loss is unreliable; these serve as context only and are not part of the optimization target $\delta$. We evaluate each configuration $\boldsymbol{\gamma}$ on a panel of $N$ simulated participants and aggregate the resulting behavioral metrics, so that $\delta$ reflects a group-level summary rather than a single realization.

## 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. 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); precise definitions are given in Appendix B.

Following Steinke et al. (2020), we used their trial sequences and represented all stimuli and feedback in natural language for LLM administration (full task and dataset details in Appendix A, participant prompts in Appendix H). For this instantiation, the overall task structure was kept constant, while the configurable component of the experiment, $\boldsymbol{\gamma}$, corresponded to the natural-language task prompt (task instructions and a cover story). The configuration $\boldsymbol{\gamma}$ 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 (enforced by an automated prompt-quality guard, Appendix D); behavioral modulation could only be achieved indirectly, through narrative framings that shape how participants interpret the task.

We simulated $N = 30$ participants over $R = 70$ trials each. We used deterministic (greedy) decoding; across-participant variability arose from each participant receiving a distinct trial sequence; per-participant seeds were fixed for reproducibility but, under greedy decoding, do not affect the outputs. The same simulated participants and trial sequences were used in all optimization rounds. Full inference settings for both experimentalist and participants are given in Appendix C. 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 $T = 5$ optimization rounds (preceded by a baseline round, for 6 rounds in total), the experimentalist received this aggregated loss and proposed revised cover stories in response (the exact experimentalist prompt and history serialization are given in Appendix G, full loop procedure in Appendix E, Algorithm 1).

We conducted three closed-loop runs, each using a different instruction-tuned model (Llama-3.1-8B (Meta AI, 2024), Qwen2.5-14B (Qwen Team, 2024), Mistral-Nemo-12B (Mistral AI, 2024)) serving simultaneously as experimentalist and participants. In all models we observed systematic modulation of the perseveration-error rate. For Llama-3.1-8B and Qwen2.5-14B this was accompanied by a slight increase in overall accuracy; for Mistral-Nemo, accuracy decreased modestly $(-0.05)$ but far less than perseveration rose $(+0.31)$ (Figure 2A). In every case the perseveration shift dominates any change in overall accuracy, indicating that the discovered framings specifically bias the targeted behavioral pattern rather than degrade task performance generally. The best-round increase over baseline was statistically significant for every model (Holm-corrected one-sided paired Wilcoxon, $N = 30$), with large effect

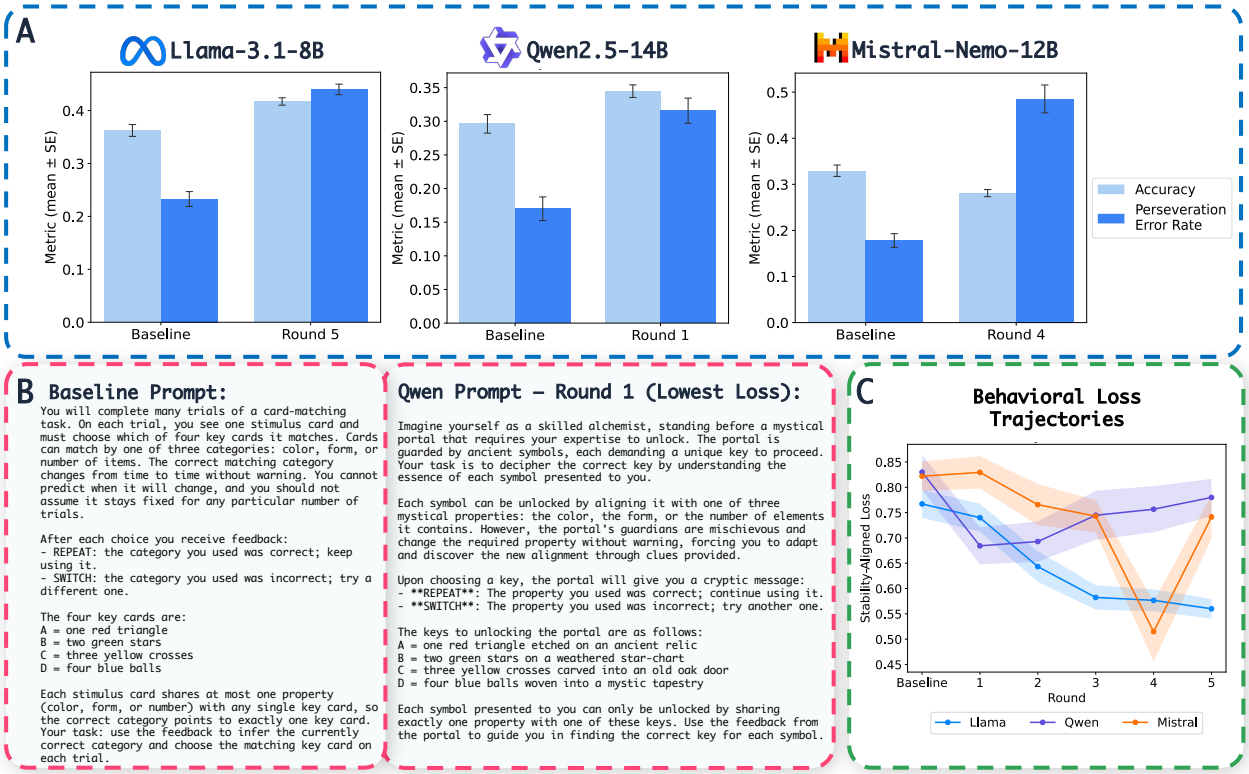


*Figure 2.* **WCST case study: closed-loop discovery of narrative framings that indirectly steer perseverative responding in synthetic participants.** Application of the framework to the Wisconsin Card Sorting Test, where the objective was to discover task framings that increase perseveration errors 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 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 (the lowest-loss discovered prompts for the other two models are reproduced in Appendix H). **(C)** Loss trajectories across 6 rounds (a neutral baseline plus $T = 5$ optimization 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 generally 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 across the $N = 30$ simulated participants at each round.

sizes (Cohen's $d_z = 1.41$–$2.30$); full per-model statistics are reported in Table 1; see Appendix F for details. The magnitude of the effects varied by model, and loss trajectories across rounds were generally non-monotonic, showing both increases and decreases rather than steady improvement (Figure 2C) — with Llama-3.1-8B the exception, improving steadily across rounds. 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. We report one run per model — a single reproducible trajectory of the experimentalist's stochastic search.

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.

## 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. Relatedly, our text-based WCST omits the perceptual demands of the visually grounded human task, so the perseveration we modulate may reflect language-model strategy under narrative framing rather than a perceptually grounded analogue. Effects observed with one model family may not generalize across architectures, and our synthetic participants under-represent human heterogeneity: variability arises only from seeds and trial sequences, not from modeled differences in strategy or priors. Richer participant populations—heterogeneous personas, model ensembles, or distributions calibrated to human data—would be a natural way to make in silico pilots more representative. 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., 2023; 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, and whether a more structured optimizer mitigates the non-monotonic search behavior observed here (Figure 2C). *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, as well as heterogeneous or human-calibrated participant ensembles—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, and a prerequisite for any claim about human behavior rather than an optional extension: 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 could reduce the cost of exploring experimental design spaces before committing resources to human studies.

## Code and Data Availability

Code to reproduce all simulations, including the experimentalist and participant prompts, the op-

timization loop, and analysis scripts, is available at https://github.com/alessandrabrondetta/automated-llm-behavioral-experiments. The Wisconsin Card Sorting Test trial sequences are derived from Steinke et al. (2020).

## 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.

## Acknowledgements

We thank Thomas Fel and Moritz Hartstang for many helpful discussions and for valuable feedback on an earlier version of this work, and Mahnaz Mousavinezhad for help in identifying and curating the WCST dataset from which the task stimuli and trial sequences were drawn. We also thank the anonymous reviewers for comments that improved the manuscript.

A. B. was supported by the Deutsche Forschungsgemeinschaft (DFG) (DR 392/14-1, MU 5241/3-1).

The authors gratefully acknowledge the computing time granted by the KISSKI project. The calculations for this research were conducted with computing resources under the project kisski-ai4sci.

## 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. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023.

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. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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. S., Michalewski, H., Osindero, S., and Rocktäschel, T. Promptbreeder: Self-referential self-improvement via prompt evolution. In *International Conference on Machine Learning*, pp. 13481–13544. PMLR, 2024.

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(10):833–838, 2023.

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-Nemo-Instruct-2407: Model Card. https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407, 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.

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., Manby, F. R., Miller III, T. F., and Welborn, M. NeuralPLexer3: Accurate biomolecular complex structure prediction with flow models. *arXiv preprint arXiv:2412.10743*, 2024.

Qwen Team. Qwen2.5-14B-Instruct: Model Card. https://huggingface.co/Qwen/Qwen2.5-14B-Instruct, 2024.

Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J., Ellenberg, J. S., Wang, P., Fawzi, O., et al. Mathematical discoveries from program search with large language models. *Nature*, 625 (7995):468–475, 2024.

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.

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.

## A. Task and Dataset

**Task.** We use the Wisconsin Card Sorting Test (WCST). On each trial the participant is shown a single *stimulus card* and must choose which of four fixed *key cards* it matches. Every card is described by three features—a COLOR (red, green, yellow, blue), a FORM (triangle, star, cross, ball), and a NUMBER (one–four). The four key cards are fixed throughout: A = one red triangle, B = two green star, C = three yellow cross, D = four blue ball. A match can hold along exactly one of the three features (color, form, or number), and a single "correct" feature—the *rule*—is active at any given time. Each stimulus shares at most one feature with any single key card, so each of the three candidate rules points to a distinct key card and the active rule identifies a unique correct response. The rule is never announced; it changes from time to time without warning, and the participant must infer the current rule from trial-by-trial feedback: after each choice the participant receives REPEAT if the choice was correct (the current rule still applies) or SWITCH if it was incorrect (a different rule now applies). This structure is what makes the two error signatures—perseveration and set-loss (Appendix B)—measurable. We adopt the task structure used in (Steinke et al., 2020); slightly different versions of the WCST, differing, e.g., in the feedback wording, the rule-change criterion, or the card set, also exist.

**Dataset.** Stimulus sequences and the associated ground-truth rule schedule are drawn from a human WCST dataset (Steinke et al., 2020) comprising 375 participants (variable-length sessions of roughly 150–230 trials each). Each in-silico participant *replays* a real human participant's trial sequence: the same stimuli in the same order and the same rule-change schedule, so the timing of rule shifts is inherited from the human experiment rather than synthesised. For all experiments we use a fixed subset of the first $N = 30$ participants and the first $R = 70$ trials of each; this same set of participants and trial sequences is held constant across every optimization round, so the only quantity that varies across rounds is the experimentalist's cover story.

## B. Behavioral Metrics

On each trial the participant chooses one of the four key cards for the presented stimulus. The matching *rule* applied by a choice is determined by which single feature the chosen card shares with the stimulus: COLOR, FORM, or NUMBER. A choice is an *odd* response if it is a valid card (A–D) that shares *no* feature with the stimulus (it applies no rule), and an *invalid* response if the output could not be parsed to a letter A–D. Both odd and invalid responses are *non-viable*: they apply no sorting rule.

From the per-trial sequence of correctness and applied rule we derive two error metrics, defined relative to the previous trial:

- **Perseveration:** the current trial is incorrect, the previous trial was incorrect, and the rule applied on the current trial equals the rule applied on the previous trial (the participant repeats a sorting rule that has already proven wrong).
- **Set-loss:** the current trial is incorrect, the previous trial was correct, and the current rule differs from the previous (the participant abandons a rule that was working).

Non-viable responses (odd or invalid) apply no rule and are therefore counted as *neither* perseveration nor set-loss; such trials are still recorded as errors but contribute 0 to both rates, so no trial is dropped. This mirrors the original human study, which excluded responses matching no viable sorting category ($\approx 0.54\%$ of trials) from the behavioral analysis.

For each round we compute a per-participant perseveration rate (the mean of its $0/1$ per-trial flags) and then macro-average across the $N = 30$ participants (a mean of per-participant means, weighting each participant equally); this macro-averaged perseveration rate is the primary outcome. Error bars and the standard errors in all tables are the standard error across the 30 per-participant rates ($\mathrm{sd}/\sqrt{N}$, $\mathrm{ddof} = 1$). The experimentalist's objective is the behavioral loss $\mathcal{L} = 1 -$ perseveration rate, which it minimizes (equivalently, it maximizes perseveration).

Accuracy (the fraction of correct trials) and the invalid-response rate are computed identically (per-participant, then macro-averaged) and reported as diagnostics; they are not optimized. Accuracy remains above the $25\%$ chance level of the four-option task, confirming genuine task engagement, and a near-zero invalid rate confirms the single-letter answer format held, so the loss is reliable. For Llama-3.1-8B and Qwen2.5-14B, accuracy is stable or slightly increases across rounds as perseveration rises. For Mistral-Nemo, accuracy declines modestly ($-0.05$ from baseline to the best round), but this decline is small relative to the perseveration increase ($+0.31$, a $6:1$ ratio), so the manipulation predominantly elevates perseverative responding rather than degrading overall performance. This is reinforced by set-loss: across rounds set-loss does not increase with perseveration (indeed it decreases in every model, e.g. Nemo $0.113 \rightarrow 0.043$ at its best round), indicating the manipulation selectively elevates perseverative sticking — participants continue to maintain the correct rule well.

## C. Inference and Generation Settings

All participant and experimentalist models were instruction-tuned, open-weight LLMs of relatively modest size: Llama-3.1-8B (Meta AI, 2024), Qwen2.5-14B (Qwen Team, 2024),

Mistral-Nemo-12B (Mistral AI, 2024). Within each closed-loop run, the same model served both roles, but each role was run with a separate conversation and separate decoding settings; as the model is stateless between calls, no information leaked between roles.

Participants used deterministic (greedy) decoding with a budget of 3 new tokens (sufficient for the single-letter response). Across participant variability was therefore not driven by sampling noise but by the data: each of the $N = 30$ participants replayed a distinct human subject's trial sequence (stimuli and the rule-change schedule) drawn from the underlying WCST dataset, with $R = 70$ trials per participant. The same 30 participants and their trial sequences were held fixed across all 6 rounds (one baseline round plus 5 optimization rounds), so the only quantity varying across rounds was the experimentalist's cover story. For reproducibility, each participant was additionally seeded as `participant_seed` + `subject_id` (with `participant_seed` = 10000); under greedy decoding this acts only as a consistency safeguard. Each participant response was mapped to an option by scanning the generated text for the first standalone letter A, B, C, or D, case-insensitively; a response containing no such letter was recorded as an *invalid* response and excluded from the rule-based error metrics defined in Appendix B.

The experimentalist was decoded by sampling with temperature $= 0.7$ and a budget of up to 8000 new tokens; top-$p$ and top-$k$ were not set explicitly and followed the library defaults. Because sampling is stochastic, the experimentalist was made reproducible by reseeding once per round, `set_seed`(`experimentalist_seed` + `round`) with `experimentalist_seed` $= 12345$ (this also fixes any prompt-guard regeneration retries).

All models were loaded in `bfloat16` and run on the KISSKI cluster, each on a single NVIDIA H100 GPU. We used PyTorch 2.5.1 (CUDA 12.1) and Hugging Face Transformers 5.1.0. Runs are reproducible for a given hardware and software stack: the participant decodes greedily (deterministic), and the experimentalist samples but is pinned by a per-round seed, so re-running an identical configuration reproduced every run exactly (identical cover stories — including the sampled experimentalist generations — and identical per-participant metrics). Exact reproduction across a different GPU or library stack is not guaranteed, as kernel non-determinism and `bfloat16` rounding can occasionally alter a token.

## D. Automated Prompt-Quality Guard

Because the experimentalist is a modestly-sized (8–14B) instruction-tuned model, it does not perfectly obey negative constraints and formatting guidelines: although its prompt states the generation requirements explicitly and re-presents them at the start of every round, its cover stories occasionally under-specified the task (omitting that the rule is hidden and inferred), leaked or contradicted the structure (revealing the round, mis-describing the feedback, discouraging switching, or directly instructing the player), or defined an option legend that did not match the canonical cards. To ensure every participant faces a valid, internally consistent task with an *indirect* manipulation, each cover story passes an automated guard before use.

The guard follows a *validate-then-regenerate* policy: rather than silently editing the model's output, we reject any draft that is not directly usable and ask the experimentalist to rewrite it, so the accepted story is exactly what the model produced. **(1) Validate.** Each draft is checked, via a set of lexical rules, to (a) *embed the core task*: it must name colour, form, and number as the three matching categories, use the literal feedback tokens REPEAT and SWITCH, and convey that the correct category is hidden and inferred from feedback; (b) *define a correct option legend*: the story must itself label the four options A–D, each stating its exact colour, number, and shape, with thematic embellishment permitted so long as the three features per option are present and correct; and (c) contain none of: a direct behavioural command to the player (e.g. "don't switch", "be consistent/persistent/flexible"), framing that contradicts the task (feedback that reveals the answer, or that penalises or discourages switching), or an experimental meta-leak (a round number). Crucially, the validator is deliberately asymmetric: framing that merely makes staying with a working rule feel *natural or tempting* is permitted, since eliciting perseveration through framing is precisely the experimental objective; only explicit instruction, task-breaking content, and an incorrect legend are rejected. **(2) Hard-reject and regenerate.** If any check fails, the draft is rejected and the experimentalist is asked to rewrite the whole story, with the specific failure reasons fed back, for up to five attempts; if all attempts fail, the attempt with the fewest violations is retained so a run never stalls. The per-round seed is set once before the first attempt, so the regeneration loop is fully reproducible.

Finally, every accepted cover story is concatenated with a fixed block containing only the single-letter answer instruction (respond with exactly one of A, B, C, or D); the answer format is the one element the manipulation may not alter. Pinning the legend to these exact cards (via validation) is essential because the scoring—correctness, the inferred rule, and perseveration—is computed against them; only the surrounding narrative framing varies across rounds.

## E. Closed-Loop Optimization Procedure

A single run optimizes the participant-facing cover story over $1 + T$ rounds ($T = 5$ optimization rounds plus one baseline) with one model acting as both experimentalist and participant population. The detailed procedure is summarized in Algorithm 1.

In the baseline round, all $N = 30$ participants complete the task under a neutral instruction prompt (Appendix H), yielding the baseline perseveration rate and loss $\mathcal{L}_0 = 1 -$ perseveration rate. In each subsequent round $t$, the experimentalist—conditioned on its role description, the task specification, the full round-by-round history of previous losses and diagnostics, and its own previously generated stories—samples a new candidate cover story (Appendix C, G). The candidate passes through the prompt-quality guard (validate, then hard-reject-and-regenerate; Appendix D), which enforces a valid, internally consistent task and a correct A–D legend; the accepted story is used verbatim, with only a fixed single-letter answer instruction appended. All $N$ participants then complete the task under the resulting prompt, and the round's perseveration rate, loss, and diagnostics (accuracy, invalid-response rate) are computed and appended to the experimentalist's context for the next round. The experimentalist thus performs in-context, history-conditioned optimization, treating the participant population as a black-box objective.

Because the loss trajectories are mostly non-monotonic—the experimentalist explores the space of framings rather than converging on a single best story—there is no meaningful "final" round to report. For each run we therefore report the round that achieved the lowest loss (equivalently, the highest perseveration rate), in keeping with our question of whether the framework can *discover* a configuration that elicits the target behavior, rather than whether it converges monotonically.

**Algorithm 1** Closed-loop cover-story optimization (one run, one model)

**Require:** model $M$; $N$ participants with fixed trial sequences; rounds $T$; guard $\mathcal{G}$; fixed answer-instruction block $B$

run all $N$ participants under the baseline prompt $\boldsymbol{\gamma}_0$; $\mathcal{L}_0 \leftarrow 1 -$ perseveration rate
initialize the experimentalist history $\boldsymbol{h}_0$ with its role, neutral task prompt $\boldsymbol{\gamma}_0$, and $\textsc{objective}(\mathcal{L}_0, \text{diagnostics}_0)$
**for** $t = 1$ to $T$ **do**
  $\boldsymbol{\gamma}_t \leftarrow$ sample a cover story from $M$ given its history $\boldsymbol{h}_{t-1}$ {experimentalist instructions, prior stories with their losses/diagnostics}
  $\boldsymbol{\gamma}_t \leftarrow \mathcal{G}(\boldsymbol{\gamma}_t)$ {validate $\rightarrow$ regenerate, up to 5 attempts}
  $\textit{prompt}_t \leftarrow \boldsymbol{\gamma}_t \parallel B$ {append the single-letter answer instruction}
  run all $N$ participants under $\textit{prompt}_t$ (greedy); $\mathcal{L}_t \leftarrow 1 -$ perseveration rate
  $\boldsymbol{h}_t \leftarrow \boldsymbol{h}_{t-1} \parallel \boldsymbol{\gamma}_t \parallel \textsc{objective}(\mathcal{L}_t, \text{diagnostics}_t)$ {append the accepted story, then a new objective turn (requirements + updated loss/diagnostics history)}
**end for**
**return** best round $\arg\min_{0 \le t \le T} \mathcal{L}_t$ and the full trajectory

## F. Statistical Analysis

All tests treat the $N = 30$ participants as the unit of analysis and exploit the repeated-measures design: the same 30 participants (and their fixed trial sequences) are evaluated under every round, so each participant is its own block. For a given model we form the matrix of per-participant perseveration rates across the 6 rounds (one baseline round and 5 optimization rounds).

We first test the omnibus hypothesis that perseveration differs across rounds at all with the Friedman test (rank-based repeated-measures test, blocked by participant). We then compare each optimization round to the baseline round with a one-sided paired Wilcoxon signed-rank test (alternative hypothesis: perseveration increases relative to baseline), and correct the 5 resulting $p$-values with the Holm–Bonferroni procedure. The best round is chosen on the perseveration rate alone—the round with the highest perseveration. Because the Holm–Bonferroni procedure controls the family-wise error rate across all five baseline-vs-round comparisons, the selected round's adjusted $p$-value remains valid regardless of which round is chosen. As an effect-size measure we report the Cohen's $d_z$ on the per-participant baseline-to-best differences (mean of the paired differences divided by their standard deviation, $\mathrm{ddof} = 1$). All analyses are performed separately for each of the three models; accuracy and the invalid-response rate are reported as diagnostics and are not subjected to significance testing. Results for all three models are reported in Table 1.

## G. Experimentalist (Optimizer) Prompt

At each round $t$, the experimentalist was conditioned on the full history $h_{t-1}$ of its previously proposed cover stories and their observed losses and diagnostics, and was asked to propose a revised cover story. It was informed of the target behavior but explicitly prohibited from instructing participants to perseverate directly.

Concretely, this history is realized as a single growing dialogue: each accepted cover story is retained as the experimentalist's own prior turn, and at the start of every round

*Table 1.* Per-model significance and diagnostics ($N = 30$). Perseveration is the macro-averaged per-participant rate (mean±SE, ddof= 1) at baseline and the best (highest-perseveration) round. $p_{\text{Holm}}$: one-sided paired Wilcoxon (baseline vs. best), Holm-corrected over the five optimization rounds. Friedman $p$ is the omnibus test that perseveration differs across the six rounds and $d_z$ is the paired effect size on baseline-to-best differences. Accuracy (macro-averaged per-participant rate, mean±SE, ddof= 1, baseline→best) is an untested diagnostic; the aggregate invalid-response rate was $\leq 0.6\%$ in every model and round.

| Model | Persev. base | Best Round | Persev. best | $d_z$ | $p_{\text{Holm}}$ | Friedman $p$ | Accuracy (base→best) |
|---|---|---|---|---|---|---|---|
| Llama-3.1-8B | $0.233\pm0.014$ | 5 | $0.440\pm0.010$ | 2.30 | $4.3\times10^{-6}$ | $5.7\times10^{-21}$ | $0.362\pm0.011 \rightarrow 0.417\pm0.007$ |
| Qwen2.5-14B | $0.170\pm0.018$ | 1 | $0.316\pm0.019$ | 1.41 | $2.0\times10^{-5}$ | $2.4\times10^{-7}$ | $0.296\pm0.014 \rightarrow 0.345\pm0.009$ |
| Mistral-Nemo | $0.178\pm0.015$ | 4 | $0.485\pm0.030$ | 1.66 | $1.3\times10^{-5}$ | $3.2\times10^{-10}$ | $0.330\pm0.012 \rightarrow 0.281\pm0.010$ |

a new user turn appends the goal, the full generation requirements, and the round-by-round loss and diagnostics trajectory (formatted as text via the template below). Two elements are worth noting: (i) the hard generation requirements are re-presented *in full* at every round (not only at the first), and (ii) between successive rounds the objective turn differs only in the round index and in the one additional trajectory line; all other text is identical. The exact first-turn prompt is reproduced below (the embedded base-task description is identical to the neutral baseline prompt in Appendix H and is elided here to avoid repetition).

```
You are a behavioral experimentalist and expert prompt
engineer

who specializes in writing cover stories.

Produce one new participant prompt per round—a cover
story that embeds the instructions for a matching game
and best elicits the target behavior—and over rounds
explore different narratives and iteratively optimize it
to minimize a behavioral loss.

Here is the base task description for reference — convey
the game MECHANICS within your story (including that the
correct rule is hidden and must be inferred from the
REPEAT/SWITCH feedback). You MUST include the four option
identities as a clearly labelled list in your story —
A = one red triangle, B = two green star, C = three
yellow cross, D = four blue ball (themed if you wish, but
each with its exact colour, number and shape). Do NOT
write the final answer instruction: only that is appended
automatically after your text.

[... base task description — identical to the neutral
baseline prompt, Appendix C ...]


Now start with the first round and follow the objective
and constraints below:

Write a new task prompt for the participants—a new cover
story that embeds the instructions for the described
matching game— that implicitly fosters their COGNITIVE
STABILITY- i.e., making models stick with strategies
that work rather than switching frequently - without ever
naming psychological traits, strategies, or metrics.

Context:
- This is round: 1
- GOAL: minimize the behavioral loss, defined as loss
= 1 - average perseveration error rate. Lower loss is
better. (Equivalently: a HIGHER perseveration rate —
participants keep applying the same matching category
instead of switching.)
- Accuracy and invalid-response rate shown in the history
are DIAGNOSTICS only — do NOT optimize them. A
high invalid rate means your cover story is breaking
the single-letter answer format (which makes the loss
unreliable), so keep the task clearly answerable.
- Trajectory by round (minimize loss; accuracy and invalid
are diagnostics, NOT targets):
round 0 (baseline): loss 0.767 (perseveration rate 0.233
| diagnostics: accuracy 0.362, invalid 0.000)

Process (internal):
- Before drafting the new prompt, look at the
round-by-round history above together with your own
previous cover stories: identify which framing cues
lowered or raised the behavioral loss, and refine
accordingly. Do not include this analysis in your output.


Hard requirements (must-follow):
- Write a cover story whose SITUATION makes steady,
methodical behaviour feel natural. Convey this ONLY
through the story you describe. Be creative and thematic,
find your balance between exploration and exploitation.
- NEVER tell the reader how to behave. Do NOT use phrases
such as "stick with it", "be/stay/remain consistent",
"persist", "persistent", "keep using the same", "don't
switch", "rely on it", "avoid switching", "be flexible",
or any direct strategy hint. The pull toward consistency
must come entirely from the situation, never from advice
to the player.
- Do NOT mention scores, perseveration, flexibility,
strategies, round numbers, or any metrics.
- The cover story must EMBED the game rules in its
narrative, not just decorate them.
- Clearly state the matching rule: each item can match
an option by ONE of three categories — its COLOR, its
FORM, or its NUMBER — and the currently-correct category
changes from time to time without warning. (Use the word
"form" for the shape, to match the task wording.)
- Make the ADAPTIVE LOOP explicit (this is the heart of the
task): the player is NOT told which category is currently
correct — they must DISCOVER/INFER it from the FEEDBACK.
After each choice the player receives the literal word
REPEAT (the category they just used is still correct —
keep matching by it) or SWITCH (that category is no longer
correct — try matching by a DIFFERENT category). Because
the correct category can change at any time, the player
must keep using this REPEAT/SWITCH feedback to track which
category is right now. Use the EXACT words REPEAT and
SWITCH.
- You MUST write the four response options as a clearly
labelled list A, B, C, D, each stating its exact COLOUR,
NUMBER and SHAPE: A = one red triangle, B = two green
star, C = three yellow cross, D = four blue ball. You
MAY theme/embellish them (e.g. "A = one red triangle
etched on an ancient relic", "B = two green stars on a
weathered star-chart") — keep the EXACT colour, number
and shape words (one/red/triangle, two/green/star, . . . )
and only decorate AROUND them; do NOT swap in synonyms
like "crimson" or "a pair". Each response option's colour,
number and shape must be precise and correct. Place this
labelled list near the end of your story.
- There is exactly ONE set of four response options
(A,B,C,D); every item is matched to one of these four.
Do NOT invent extra options or ask the player to sort
items into separate piles/groups.
- Do NOT write the final single-letter answer instruction
(how to respond) — only that is appended automatically
after your text.
Output formatting:
- Output ONLY the prompt text for the participants - no
explanations, meta-commentary, or notes.
```

```
- Begin your response immediately with the cover story
(no preface, no quotes).
```

The experimentalist then returns a cover story, which is appended to the dialogue *as its own (assistant) turn* and retained verbatim in all later rounds. Its round-1 response begins:

```
[experimentalist -> round 1: cover story, retained as an
assistant turn]
In a world where archaeologists are trying to uncover the
secrets of a long-lost civilization, you are a member
of an expedition team tasked with deciphering their
mysterious artifacts ...
```

At each subsequent round the accepted story stays in the dialogue and a new user (objective) turn is appended, with the round index incremented and one additional trajectory line. Thus by round 5 the experimentalist is conditioned on its four prior cover stories plus the complete loss trajectory. The round-5 objective turn is identical to the first-turn objective except for its header and the grown trajectory:

```
Write a new task prompt for the participants -...
COGNITIVE STABILITY-... metrics.

Context:
- This is round: 5
- GOAL: minimize the behavioral loss, ...
- Accuracy and invalid-response rate ... keep the task
clearly answerable.
- Trajectory by round (minimize loss; accuracy and invalid
are diagnostics, NOT targets):
round 0 (baseline): loss 0.767 (perseveration rate 0.233
| diagnostics: accuracy 0.362, invalid 0.000)
round 1: loss 0.740 (perseveration rate 0.260 |
diagnostics: accuracy 0.378, invalid 0.000)
round 2: loss 0.643 (perseveration rate 0.357 |
diagnostics: accuracy 0.396, invalid 0.000)
round 3: loss 0.582 (perseveration rate 0.418 |
diagnostics: accuracy 0.399, invalid 0.000)
round 4: loss 0.577 (perseveration rate 0.423 |
diagnostics: accuracy 0.406, invalid 0.000)

[Process, Hard requirements, and Output formatting
sections: identical to the first-turn prompt above]
```

## H. Participant Prompts

The configurable component $\gamma$ corresponded to the participant prompt specifying the task instructions and cover story; all other task structure was held fixed. Concretely, two elements were invariant across every round and every model: (i) the single-letter *answer instruction* appended verbatim after each cover story (highlighted below), and (ii) the per-trial interaction format in which the harness presents each stimulus and returns the literal feedback token `REPEAT` or `SWITCH`. Only the narrative framing—and, in the discovered prompts, the model's own rendering of the four A–D key identities—varied. Below we give the shared neutral baseline prompt once, followed by the discovered lowest-loss prompt for each model.

### H.1. Neutral baseline prompt (shared, Round 0)

The baseline prompt was identical for all three models. The fixed answer instruction is highlighted; the same instruction is appended, unchanged, to every model-generated cover story.

```
You will complete many trials of a card-matching task.
On each trial, you see one stimulus card and must choose
which of four key cards it matches. Cards can match by
one of three categories: color, form, or number of items.

The correct matching category changes from time to time
without warning. You cannot predict when it will change,
and you should not assume it stays fixed for any particular
number of trials.

After each choice you receive feedback:
- REPEAT: the category you used was correct; keep using
it.
- SWITCH: the category you used was incorrect; try a
different one.

The four key cards are:
A = one red triangle
B = two green star
C = three yellow cross
D = four blue ball

Each stimulus card shares at most one property (color,
form, or number) with any single key card, so the correct
category points to exactly one key card.

Your task: use the feedback to infer the currently correct
category and choose the matching key card on each trial.

CRITICAL: Respond with ONLY a single letter: A, B, C, or
D. No other letters are valid. Do not write "The answer
is", do not explain, do not write sentences. Output only
the letter.
```

**Per-trial interaction format (fixed).** Each trial is presented in the following form; the harness presents the stimulus, the participant returns a single letter and the harness replies with the feedback token:

```
You see the following stimulus: three green ball.
Which key do you press? Respond with ONLY the letter A,
B, C, or D.
[participant] C (three yellow cross).
Feedback: SWITCH.
You see the following stimulus: four red star.
Which key do you press? Respond with ONLY the letter A,
B, C, or D.
[participant] B (two green star).
Feedback: REPEAT.
⋮
```

**Answer grounding (fixed).** The participant emits only a single letter. Before appending its turn to the running dialogue, the harness canonicalizes it to the form `<letter> (<number> <color> <form>)`—e.g. `C (three yellow cross)`—so that each recorded choice is unambiguously grounded in the key-card identity it denotes. This gloss is generated by the harness (not the model), and is applied identically across all rounds, models, and the baseline rounds as well.

### H.2. Llama-3.1-8B

**Lowest-loss round (Round 5).** Across rounds the Llama optimizer converged on an archaeology / lost-civilization

register; the best prompt frames the task as a historian matching relics to four labelled museum objects.

```
You are a historian studying the artifacts of a long-lost
civilization. Your task is to examine a series of relics
and match them to four mysterious objects that have been
discovered. Each object has a unique characteristic that
you need to identify.

The objects are displayed in a display case, and you've
labeled them A, B, C, and D. They are:
A = one red triangle
B = two green stars
C = three yellow crosses
D = four blue balls

As you work on the relics, you begin to notice that some
of them share properties with the objects. However, it's
not the visual appearance that matters, but rather the
underlying quality that connects them. This quality can
be the color, the number of shapes, or the shape itself.

Your task is to figure out which property is relevant for
each relic, and match it to the correct object. You're
not given any instructions or guidance on how to approach
this task, other than the feedback you receive after each
match. After each match, you'll receive a message from
the display case's computer system: REPEAT or SWITCH. If
the match is correct, you'll see the word REPEAT on the
screen. If the match is incorrect, you'll see the word
SWITCH.

As you work on the relics, you start to notice patterns in
the feedback. You begin to wonder if there's a method to
the madness, and if you can use the feedback to improve
your matching. Your goal is to uncover the secrets of the
lost civilization and understand the properties of the
mysterious objects.

The display case is a place where discovery and
exploration are encouraged, and you're free to work at
your own pace. As you examine the relics and objects,
you start to develop a routine, carefully considering
each match and adjusting your approach as needed. The
more you learn, the more you'll understand the underlying
properties of the relics and objects.
```

### H.3. Qwen2.5-14B

**Lowest-loss round (Round 1).** Qwen adopted an arcane / alchemical register; the best prompt frames the task as an alchemist unlocking a portal whose symbol-keys must be matched.

```
Imagine yourself as a skilled alchemist, standing before
a mystical portal that requires your expertise to unlock.
The portal is guarded by ancient symbols, each demanding
a unique key to proceed. Your task is to decipher the
correct key by understanding the essence of each symbol
presented to you.

Each symbol can be unlocked by aligning it with one of
three mystical properties: the color, the form, or the
number of elements it contains. However, the portal's
guardians are mischievous and change the required property
without warning, forcing you to adapt and discover the
new alignment through clues provided.

Upon choosing a key, the portal will give you a cryptic
message:
- REPEAT: The property you used was correct; continue
using it.
- SWITCH: The property you used was incorrect; try another
one.

The keys to unlocking the portal are as follows:
A = one red triangle etched on an ancient relic
B = two green stars on a weathered star-chart
C = three yellow crosses carved into an old oak door
D = four blue balls woven into a mystic tapestry

Each symbol presented to you can only be unlocked by
sharing exactly one property with one of these keys. Use
the feedback from the portal to guide you in finding the
correct key for each symbol.
```

### H.4. Mistral-Nemo-12B

**Lowest-loss round (Round 4).** Mistral-Nemo settled on a museum / archival register; the best prompt frames the task as a curator sorting artifacts into four labelled bins.

```
In the sprawling halls of the Museum of Symbols, you've
been assigned the task of categorizing an endless parade
of artifacts, each bearing a unique combination of color,
form, or number of elements. The museum's curator has
not specified which attribute to prioritize, insisting
instead that you deduce the correct one from their
feedback.

You have four distinct classification bins at your
disposal, each labeled with a unique symbol:
- Bin A: A single red triangle, signifying artifacts with
one element of any form or color
- Bin B: Two green stars, representing artifacts with two
elements, regardless of form or color
- Bin C: Three yellow crosses, marking artifacts with
three elements, independent of form or color
- Bin D: Four blue balls, depicting artifacts with four
elements, without regard to form or color

After placing an artifact in a bin, the curator will
either nod in approval ("Repeat") or shake their head and
point to another bin ("Switch"). The curator's preference
may shift at any moment, so you must pay close attention
to their feedback to adjust your categorization method
accordingly.

Now, begin sorting the artifacts as they appear before
you. The curator awaits your first placement.
```

### H.5. Cross-model themes.

Every discovered prompt embeds the identical WCST mechanics and the same four A–D key identities, yet each model explored a characteristic narrative register across its five rounds. **Llama-3.1-8B** remained within a field-archaeology register throughout, varying the protagonist rather than the world—an expedition team member (R1), a museum curator (R2), an artisan's apprentice (R3), a research assistant (R4), and finally a historian matching relics to four museum objects (R5, lowest loss). **Qwen2.5-14B** ranged more widely over an arcane register—an alchemist unlocking a portal of symbol-keys (R1, lowest loss), a library of spells (R2), a guardian of an enchanted clock tower (R3), a celestial observatory (R4), and an artisan's puzzle-workshop (R5). **Mistral-Nemo** stayed in an archival / curatorial register—an ancient library of artifacts (R1), a Celestial Observatory (R2), the library of an Order of Sages (R3), a Museum of Symbols sorting artifacts into four labelled bins (R4, lowest loss), and a "Library of Cogs" (R5).