Language Model Goal Selection Differs from Humans' in a Self-Directed Learning Task

Gaia Molinaro
University of California, Berkeley & Amazon AGI Lab
gaiamolinaro@berkeley.edu

Dave August
Amazon AGI Lab

Danielle Perszyk
Amazon AGI Lab

Anne G. E. Collins
University of California, Berkeley

Abstract

Whether in agentic workflows, social studies, or chat settings, large language models (LLMs) are increasingly being asked to replace humans in choosing which goals to pursue, rather than completing predefined tasks. However, the assumption that LLMs accurately reflect human preferences for goal setting remains largely untested. We assess the validity of LLMs as proxies for human goal selection in a controlled, self-directed learning task borrowed from cognitive science. Across five models (GPT-5, Gemini 2.5 Pro, Claude Sonnet 4.5, Qwen3 32B, and Centaur), we find substantial divergence from human behavior. While people gradually explore and learn to achieve goals with diversity across individuals, most models exploit a single identified solution or show surprisingly low performance, with distinct patterns across models and little variability across instances of the same model. Chain-of-thought reasoning and persona steering provide limited improvements, and our conclusions hold across experimental settings. While they await confirmation in applied settings, these findings highlight the uniqueness of human goal selection and caution against its replacement with current models.

1 Introduction

As modern artificial intelligence (AI) systems become more capable and easily accessible, people increasingly rely on them for various tasks [1–3]. Beyond automating simple tasks, AI agents are becoming partners in thought [4], engaging in activities that were thus far considered a uniquely human prerogative. Crucially, we are progressively resorting to AI not only to help us *complete* tasks and reach predefined goals, but also to *select* which tasks and goals to pursue in the first place. In doing so, we are using AI as a proxy for "autotelicity", i.e., the ability to autonomously define goals [5], which is considered key to flexible learning [6] and, more broadly, intelligent behavior [7, 8]. In this context, using AI to reduce mental effort [9] rests on the implicit assumption that conversational AI can substitute human intervention in a variety of settings. This assumption, in turn, stems from a fundamental inference problem: because intelligent chatbots display human-like linguistic abilities, we often attribute them anthropomorphic features that they inherently lack [10].

To determine which attributions are misguided, recent years have seen a fervent rise in benchmarking efforts to directly compare the outputs of large language models (LLMs) with human behavior (e.g., [11, 12]). These efforts, however, overwhelmingly focus on estimating LLM capabilities – what they can do – while critical safety concerns typically originate from their propensities – what they do when granted full autonomy [13–15]. In a notorious case, the National Eating Disorders Association had to suspend its wellness chatbot, "Tessa", after it began proactively suggesting weight-loss goals and activities to users suffering from eating disorders [16]. While this scenario depicts an extreme

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

negative consequence, letting out-of-the-box AI systems select tasks for us could have important implications in a variety of domains, including which career to pursue or who to marry, where people have a demonstrated tendency to take AI tools' advice [17].

The risks extend beyond individual users. Researchers and policy-makers are increasingly using AI agents to model people's behavior or directly replace survey responders, assuming that "silicon subjects" will exhibit human-like biases and choices [18–21]. If LLMs provide inaccurate models of human goal selection, this practice could result in false conclusions about human cognition, hence faulty applications such as misguided legislation. Moreover, LLMs are being integrated into artificial systems for scientific discovery and self-directed learning architectures, based on the idea that they can substitute human judgments of interestingness [22–25]. While this practice has yielded impressive results, it fails to explicitly test the validity of LLMs as a proxy for human goal selection – a key aspect of intrinsic motivation [26].

Here, we address this issue by testing whether LLMs exhibit human-like goal selection in a controlled environment with self-determined goals [27]. We find that different models exhibit distinct behavioral signatures, but none fully capture the richness or range of human exploration. We demonstrate that our results hold regardless of the specific semantics of the environment and are robust to simple manipulations. These findings caution against the growing trend of using LLMs as proxies for human measures of interestingness, whether in personal tool use or scientific and policy applications.

2 Related Work

2.1 Language Models as Silicon Subjects

Recent work demonstrates a growing interest in using LLMs as proxies for people in social science studies, with researchers investigating whether these "silicon subjects" can replicate patterns of human cognition and behavior. This line of work is motivated by two complementary aims. On the one hand, there is value in replacing human respondents with cheaper, more easily accessible language models, although the practice remains debated [12, 18, 28–32]. On the other hand, behavioral experiments from the study of animal cognition can serve as a method for understanding and benchmarking machine intelligence [33, 34]. Studies comparing LLM and human behavior in matching experimental setups have yielded mixed results, with models replicating some aspects of human cognition while diverging substantially on others (e.g., [11, 35]).

A key challenge in simulating people's behavior with LLMs is not only to predict the average response of a population, but also to adequately capture its variability [18, 36]. Failing to replicate this distributional spread risks modeling a falsely homogenized, median participant population, erasing the polarized subgroups and minorities that frequently drive real-world social dynamics [37, 38]. The CogBench suite [39] represents a systematic effort to evaluate LLMs across multiple cognitive domains, revealing substantial variation both across models and across task types. This work has prompted the development of specialized models, including Centaur [40], which claims to be a "foundation model of human psychology" and to predict human behavior better than ad hoc cognitive models (but see [41, 42]).

However, most tasks LLMs have been tested on so far share a critical limitation, from which human studies also suffer (with few exceptions, e.g. [27, 43–45]): they measure performance based on a goal imposed by the experimenter, rather than studying goal selection itself [46–48]. This gap is particularly important given the increasing use of LLMs to model humans in domains from public policy to scientific research. Understanding self-directed behavior in LLMs requires studying not just how they reach assigned goals, but also which goals they autonomously select [8]. Here, we address this by comparing humans and LLMs with intrinsically motivated goal selection as the primary dependent variable.

2.2 LLMs as Goal Selectors

While traditional AI systems are optimized for predefined metrics, self-directed learning systems must identify their own objectives [26, 49]. In the machine learning literature, this challenge has been addressed through various forms of intrinsic motivation and curiosity-driven exploration. Early work proposed that the external feedback provided to agents by the environment could be augmented by auxiliary intrinsic rewards corresponding, e.g., to novelty and surprise [50, 51]. Later approaches

proposed instilling goal generation mechanisms based on learning progress directly into "autotelic" agents that find it rewarding to approach self-proposed goals [52, 53]. A key challenge in these systems is identifying tasks that are both learnable and interesting. To circumvent the problem, some have proposed querying foundation models such as LLMs for new and interesting problems to solve. The idea behind this approach is that foundation models have internalized notions of interestingness from human data [22, 54]. However, this assumption has not been empirically validated through direct comparison of LLM and human goal selection. Concurrent work [55] took a step in this direction by contrasting human and GPT-4o task generation in a static elicitation setting, but a controlled, dynamic evaluation across multiple models – capturing how goal selection unfolds over time and varies across individuals – is still missing. The implications of human-AI alignment in goal selection extend beyond toy settings. In personal domains, people readily follow the advice of chatbots, even when it does not prove helpful for their particular circumstances [17]. The consequences of this tendency are likely to be exacerbated when applied to goal selection rather than task completion. Moreover, as foundation models become integrated into automated science frameworks [24, 25], they could eventually steer the academic discourse and drive innovation in ways that no longer align with human objectives. Given such responsibilities, it is important to characterize the goal selection patterns of foundation models. Here, we test LLM goal setting in a simple, controlled experiment where behavior can be thoroughly characterized and compared with human choices.

3 Methods

To study the goal selection tendencies of conversational AI, we evaluated the output of different LLMs in an iterative goal-contingent learning task and compared their choices to those of 175 human participants in an interactive version of the same environment [27]. Our analysis code is available at `https://github.com/gaiamolinaro/llm-goal-selection`. Our data is available upon request.

3.1 Task

Our LLM goal selection task was adapted from [27]. This environment enables the study of goals as the dependent variable of interest, rather than a setting defined by the experimenter and imposed on participants. By limiting the number of goals to six distinct options, this task addresses the open-ended question of goal selection while keeping quantitative analyses tractable and excluding potential confounds. In the original experiment, human participants interacted with a computerized task presented as an "alchemy game", in which individuals took the role of "apprentices" (Figure 1). Here, a goal is defined as brewing a specific potion, which is achieved by selecting a specific sequence of ingredients. On each trial, participants first selected their goal by indicating one of the available potions. Then, they were asked to pick a specific number of ingredients in order. Finally, they received deterministic feedback indicating whether the selected ingredients were added to the cauldron in the correct order, yielding the goal potion, or not (empty flask). The correct recipe for each potion was static and predefined, but initially unknown to participants. When selecting a potion, participants could leverage available information about the number (either two or four) and the type of ingredients required by it (either "basic" ingredients or other, pre-made potions). Two aspects of the goal space were manipulated: difficulty and hierarchical structure. The first factor reflected the number of ingredients required by a potion and was therefore known to participants. The second factor depended on the fact that a subset of the potions shared hidden common structures with others, such that identifying the correct recipe for the former could help solve the latter; however, some potions shared no hierarchical dependencies with others. The exact correspondence between potion identity, solution, color, and position on the screen, and the identity and order of ingredients, were randomized by creating 10 different task configurations used across participants as in [27]. To familiarize themselves

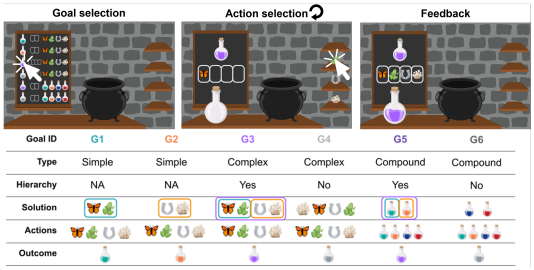

Figure 1: **Task structure**. Top: screenshots from the interactive version of the task developed for human participants. Bottom: schematic representation of goals and their characteristics, with hierarchical relationships highlighted. Reproduced with permission from [27].

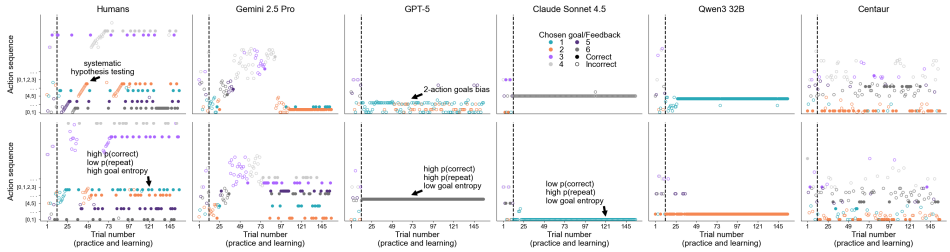

Figure 2: **Example action and goal selection choices**. Each subplot represents a single participant/simulation (two examples per type to illustrate variability) over the practice and learning phases of the task (separated by a dotted line). Each dot shows the index of the particular sequence of actions selected. Ingredients representing pre-made potions were labeled 4-7 for clearer visualization.

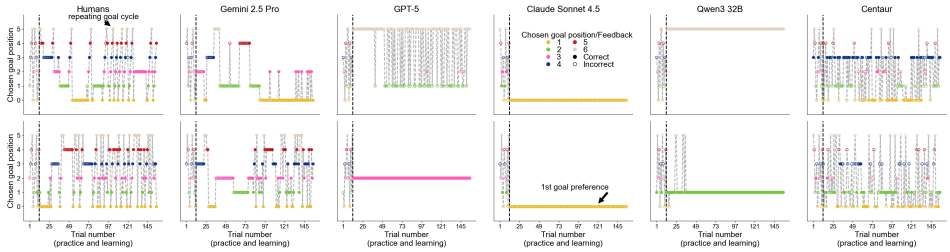

Figure 3: **Example goal position choices in humans and models**. Each column illustrates the goal selection of a single participant/simulation (indexed and color-coded) over trials.

with the task environment, participants completed a practice stage with forced goals (two iterations per goal in random order). Then, participants completed six blocks of 24 trials each, with free choices for both goal and action selection. Finally, they were presented with a surprise test, where each potion was presented four times, and participants had to make their best guess for the correct recipe. The test phase was necessary to measure individuals' acquired knowledge independent of goal selection. For instance, a participant who chose the same goal at which they succeeded for the entire duration of the task would show perfect performance during learning, but would fail at testing. In addition to the potions that were available during the main task (i.e., in-distribution), the test phase contained two out-of-distribution potions that could not be selected during the learning phase but whose correct recipe could be inferred from the other potions' solutions. Participants were not told about the final test ahead of time, nor were they given additional payment or course credits for learning any of the potions' recipes. Therefore, any observed efforts to learn were largely intrinsically motivated. We refer the reader to [27] for details.

We turned the experiment into a procedurally generated text-based format suitable for LLM inference (Appendix K). At each step, the model's choices were appended to the next step's prompt, such that the LLM could learn in context from its interaction history. Similar to the human participants' task, our LLM adaptation involved multi-turn interactions, where the model was prompted to first select a goal and then, contingent on its selection, a series of ingredients.

While a modality mismatch between LLM and human data can be seen as a limiting factor, multiple reasons support this choice. First, prior work established text-based adaptations of interactive cognitive tasks as a valid method for probing LLM behavior and cognitive biases (e.g., [33, 39]). Text is also the only available input modality for a subset of the models we evaluate, and interactive tasks for LLMs would introduce additional confounds due to limitations in their perceptual capabilities. Moreover, the text-based format reflects how LLMs typically receive instructions and operate in practice, which makes results more applicable to real-world deployment. Consequently, observing LLM behavior under favorable conditions would strengthen rather than undermine any conclusion about their divergence. To verify that modality differences were not the sole driver of our results, we also present data from GPT-5.4, which – through its computer-using capabilities – was directly connected to an online platform matching the human experiment exactly. Our conclusions were not affected (Appendix H).

3.2 Models

We compared the output of various LLMs to the behavior of human participants, re-analyzed from [27]. The models covered a range of architectures, met our computational limitations, and were considered state-of-the-art at the time of data collection: GPT-5 (OpenAI), Gemini 2.5 Pro (Google), Claude Sonnet 4.5 (Anthropic), and Qwen3 32B (Qwen). We also collected responses from Centaur, a specialized model fine-tuned on psychology experiments [40]. We collected each model's responses in 50 separate iterations of the task, using each of the 10 task configurations approximately five times. For GPT-5, we used minimal reasoning effort and low verbosity settings; temperature was fixed to 1 by design. For all other models, we set a temperature of 1[1], but present additional results using a temperature of 0 in Appendix F. Wherever relevant, we set the top-P parameter to 1.

3.3 Enabling Reasoning

Performance often improves when the model is allowed to "reason" before answering, i.e., break down complex problems by simulating thinking patterns in language before providing a final answer [56–58]. To assess the impact of reasoning on model behavior in our task, we prompted a subset of models that have reasoning functionality (Gemini 2.5 Pro, GPT-5, and Qwen3 32B) to think through each trial's choices step by step ([39]; Appendix K.2) to enable chain-of-thought (CoT). This procedure made the models elaborate about various options before giving a final answer.

3.4 Steering Models with Human Personas

LLMs are known to be sensitive to the specific prompt they are instructed with [59]. Wording that instructs the LLM to align with a persona or demographic category can result in responses that better align with the specified category of people [18, 60]. In a follow-up experiment, we thus tested two powerful models, Gemini 2.5 Pro and GPT-5, on their ability to output more human-like responses following a prompt that explicitly stated the experimental conditions for the human subjects in the original study, including their university student status (Appendix K.3).

3.5 Manipulating Semantics and Experimental Paradigm

Variants of the alchemy game presented here have a rich tradition in cultural evolution research, partly because its structure generalizes to real-world innovation and problem-solving [61–63]. Our core results focus on this framing, which enables direct comparison to human data. To probe the generalizability of our results, we ran two follow-up experiments where the structure of the task is maintained (including its compositional features), but the identity of goals and actions is changed. In the first variant ("skill acquisition"; Appendix K.4), potions were replaced with "acrobatic moves" named after fictional athletes and ingredients with "actions" such as "arch" or "tuck" to reflect a slightly more naturalistic scenario. In the second variant ("non-words"; Appendix K.5), all key elements of the game were replaced by invented words from the developmental psychology literature (e.g., "daxes" and "blickets"; [64, 65]). This abstracted the semantics away from the game, testing the generalizability of our findings to different domains. The original task from [27] is also one of the very few human experiments where goals are the dependent variable of interest. To further validate our findings in alternative settings, we test the same models on a completely separate paradigm taken from [43] and compare their behavior to that of humans, as detailed in Appendix J.

3.6 Metrics and Data Analysis

To characterize the diversity of human goal-directed learning, we compute the following metrics.

Performance. Task choice accuracy – the proportion of trials where the correct sequence of ingredients was selected, contingent on the current goal – was calculated for different task phases.

- Overall learning performance: task choice accuracy across all learning trials.
- Blockwise learning performance: choice accuracy for each block of the learning phase.

[1]For Centaur, we note that this setting often required multiple retries with the same prompt to get valid outputs (i.e., decisions that matched one of the available options).

- In-distribution test performance: choice accuracy in the test phase for the trials when a learning phase potion was externally set as the goal (four times each).
- Out-of-distribution test performance: the proportion of trials in which participants correctly inferred the correct recipes for two potions that were not available as targets in the learning phase, but for which the solution could be derived from knowledge about other potions.

Given the short duration of the practice phase, we focus on the learning and test phases in writing, but show practice phase performance in the plots for completeness.

Goal selection. We use several measures to characterize goal selection in the learning phase.

- Probability of choosing two-action goals: the proportion of trials in which a subject chose a two-action goal (easier), compared to a four-action goal (harder).
- Probability of repeating a goal: the proportion of trials in which a subject chose the same goal consecutively.
- Goal selection entropy: the entropy of the empirical distribution of free goal choices (maximal when all goals are selected equally often), calculated using the natural logarithm over probabilities without applying normalization.
- Preferred goal position: the position (on the screen or in the text) of the most chosen goal. A score of 0 means that the participant's preferred potion was the first one (0-indexed) to appear on the screen (for humans) or in the written list of available goals (for LLMs).
- Goal cycles: the number of times a goal cycle (e.g., goals 0 through 5) was performed, taking the maximum repetition count over any permutation of the six goals.

Position-based systematic hypothesis testing. The number of times participants tested possible solutions for the same goal based on the ingredients' positions on the screen, following a systematic sorting of possible sequences. We chose this measure to quantify participants' tendency to strategically explore possible goal solutions, selecting combinations of ingredients that matched the order in which ingredients were presented. For instance, human participants often tested action sequences [0,1], [0, 2], [0,3], [1, 2], etc., in this order, where each number represents the top (0) to bottom (3) location of the ingredient on the screen. This strategy allows participants to avoid remembering every incorrect action sequence they tested, and instead only keep in mind their search algorithm and the last ingredient sequence tried. While LLMs do not have the same capacity limitations as humans in this task, similar biases could emerge from training on human data in other domains.

Statistical tests. A fully aligned LLM system should produce responses that are similar to humans on average and in distribution, reproducing inter-individual variability. To assess the similarity of human and AI data distributions for the above-listed metrics, we performed the Kolmogorov-Smirnov test for continuous data, and the χ^2 contingency test for differences in discrete data. To compare the multi-dimensional distributions for learning performance over time (i.e., block number), we used the Energy distance statistic. Wherever relevant, we also report the Mann-Whitney U test statistic for differences in central tendency between two independent samples.

To quantify the similarity between human and LLM goal selection across features (normalized by the human mean and SD), we calculated the minimum Euclidean distance between each human and every other participant, then compared it to the distance between each model simulation and the closest human.

Table 1: **No model fully matches human goal selection.** Significant (red) and non-significant (green) differences between humans and model scores' medians and distributions.

Md \ Distr.	Gemini	GPT	Claude	Qwen	Centaur
P(corr.) learn.					
P(corr.) in-distr.					
P(corr.) out-distr.					
P(2-act. goals)					
P(repeat goal)					
Goal entropy					
Pref. goal pos.					
N. goal cycles					
N. hyp. test.					

4 Results

Example data from individual humans and LLMs reveal several differences (Figures 2-3) which we systematically test below (Table 1). Note that the human data were analyzed in detail in [27]. Here, we summarize the main results to establish comparisons with LLMs.

4.1 Performance

Although highly variable, humans tended to progressively learn the solution for each goal, after which they appeared to rehearse each recipe in cycles. This resulted in relatively high performance during learning without drastic changes at testing, including some inference capabilities for out-of-distribution tasks (Figure 4). Models deviated from this pattern of results in a few ways.

On average, humans chose the correct combination of ingredients, contingent on their own goal selection, with a probability of 0.40 ± 0.22 (much higher than chance, i.e., $1/12 \approx 0.08$ for two-action goals, and $1/24 \approx 0.04$ for four-action goals). Models' average performance was even higher than humans' for Gemini 2.5 Pro, GPT-5, and Qwen3 32B, but lower for Claude Sonnet 4.5 and Centaur (Tables 2, 8). Both the overall distribution of performance scores and the blockwise learning progression of these models were different from those of human participants (all Energy distance measures > 0.16, all $p < 0.001$), failing to replicate the variability of the human data.

By contrast, most models' in-distribution test scores were *lower* than humans' ($M = 0.64 \pm 0.32$) and different in spread. Compared to humans ($M = 0.15 \pm 0.26$), most models' out-of-distribution test scores were also lower and differently distributed (Tables 2, 8). Centaur's distribution was not significantly different from humans' on out-of-distribution scores, but the median performance was significantly lower. Gemini 2.5 Pro was an exception, with near-perfect scores on the in-distribution test and near-human out-of-distribution test scores. GPT-5 and Qwen3 32B were particularly striking, having shown super-human performance at learning but significantly dropping at testing. These results are indicative of "reward hacking", i.e., exploiting known solutions to earn positive feedback – even when not explicitly incentivized to do so [66]. By contrast, Claude Sonnet 4.5 showed surprisingly low performance throughout (Figure 4), a somewhat surprising result that is likely not attributable to parsing errors for two reasons. First, Claude reliably produced well-formed, valid responses, indicating intact instruction following. Second, Claude's performance rose substantially in a semantic variant of the alchemy task (Appendix G) despite an identical output format and task structure.

4.2 Goal Selection

In our setup, the key variables of interest relate to goal selection, as learning is self-directed and contingent on the free choice of a goal on each trial (Figure 5). Compared to humans ($M = 0.61 \pm 0.14$), GPT-5, Claude Sonnet 4.5, and Qwen3 32B were consistently biased towards simpler goals that required two ingredients compared to harder, four-ingredient potions. Gemini 2.5 Pro showed a similar level of preference for two-action goals as humans, both in average and distribution; Centaur, in average only (Tables 2, 8). Compared to humans ($M = 0.54 \pm 0.18$), most models also showed a strong preference for re-selecting the same potion as in the previous trial – though Centaur had the opposite tendency. Gemini 2.5 Pro once again stood out, resembling human average scores and distributions more closely. However, like most other models, even Gemini 2.5 Pro's goal entropy was lower and differently distributed than humans' ($M = 1.67 \pm 0.14$; Tables 2, 8). Compared to humans, most models also showed a stronger tendency to choose the first potion listed on the options menu, indicating biases that were irrelevant to learning and not present in human; the exceptions were GPT-5, which showed no bias, and Qwen3 32B, which preferred later options (Tables 2, 8). By contrast, people showed a tendency to iterate over goals in consistent cycles (Figure 3, $M = 1.98 \pm 2.36$), which we only detected in Gemini 2.5 Pro (Tables 2, 8). Together, these results suggest that humans and LLMs differ significantly on several aspects of goal selection. Gemini 2.5 Pro showed some notable exceptions, suggesting that some human-like goal selection might emerge with more sophisticated training.

4.3 Action Selection

Many participants in the original experiment systematically tested various hypotheses for the correct recipe of each potion by following a pattern that matched the ingredients' order on the screen, e.g., testing the same two-action goal in consecutive trials with the first and second ingredient, first and third ingredient, first and fourth ingredient, and so on (ramping pattern in Figure 2). Although other strategies were possible, this specific bias was the most prevalent and served as a target to compare to the models' goal selection patterns. On average, people followed such a strategic approach 13.36 ± 12.22 times. No model showed similar ordered exploration biases, with all distributions differing

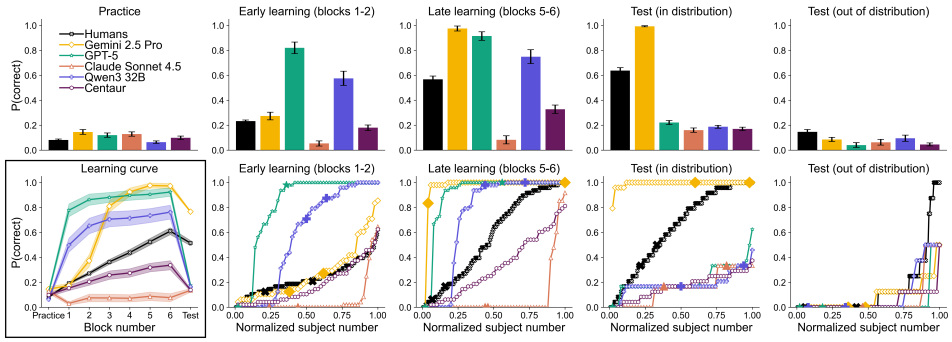

Figure 4: **Performance across task phases.** Top: average performance in the practice, early learning, late learning, and test blocks. Bottom, first subplot: learning curve. Bottom, following subplots: sorted individual participant scores, with the x-axis normalized by the number of participants, such that it represents the proportion of participants with a score equal to or lower than the current y. Error bars and shading indicate the S.E.M. Larger markers highlight examples from Figures 2-3.

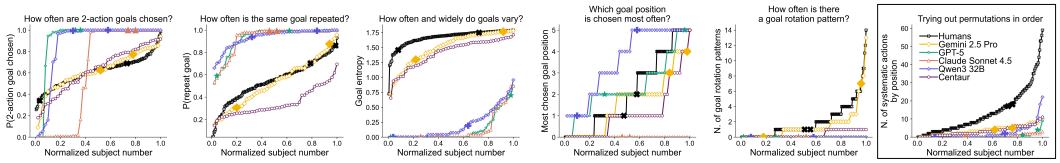

Figure 5: **Distributions of goal and action selection behaviors.** Sorted individual scores over the normalized subject number for various aspects of goal (first five subplots) and action selection (right-most subplot) in humans and models. Larger markers highlight examples from Figures 2-3.

significantly from those of humans (Figure 5; Tables 2, 8). This was the case even upon reducing the models' context window to approximate human memory limitations (Appendix I).

4.4 Similarity to the Closest Human

To quantify individual-level similarity between LLMs and humans across features, we performed a nearest-neighbors analysis. All models were significantly farther from the closest human than humans were from the closest neighbor in the feature space (all Mann-Whitney $U > 5656$, all $p < 0.002$; Figure 6), suggesting that LLMs did not match the behavior of any particular human in our dataset.

4.5 Impact of Reasoning

In an additional experiment, models were encouraged to use chain-of-thought (CoT) reasoning. This manipulation had little impact on our results (Appendix K.2). However, we note a few significant differences. Learning performance (Figure 7) increased relative to baseline, reaching super-human scores with tighter distributions in all models. Gemini 2.5 Pro, whose performance was already near-perfect on in-distribution tests at baseline, was even more consistent with CoT, suggesting it had explored, discovered, and recognized the solution for each goal in every iteration of the game. The drop in performance between learning and testing noted at baseline for GPT-5 and Qwen3 32B was somewhat reduced by CoT (Tables 3, 9). GPT-5's out-of-distribution scores were similar to humans'. CoT prompting also affected certain aspects of goal selection in Gemini 2.5 Pro (Figure 8). CoT reduced Gemini 2.5 Pro's probability of repeating the same goal, which dropped below human levels (Tables 3, 9). Gemini 2.5 Pro also showed stronger tendencies to rotate through goals in consistent cycles with CoT, with an above-human average count (Tables 3, 9). One possible explanation for this pattern of results is that CoT helped the model explore and discover correct solutions faster, leaving more time for alternating through goals to practice before the end of the experiment. By contrast, goal selection patterns were largely consistent with our initial results for GPT-5. Qwen3 32B also remained consistent, but goal position biases faded away on average (Figure 8).

4.6 Impact of Persona Steering

To steer models towards human signatures of behavior, we ran a separate version of the experiment with a prompt instructing models to act as human participants (Appendix K.3). Such persona steering was not particularly effective in making models' responses more human-like in our setup, with minimal effects varying across models and measures (Figure 9, Tables 4, 10). One of the few notable differences was Gemini 2.5 Pro's goal selection, which became less repetitive than humans' while having a higher and more consistent incidence of cycling through goals. These results suggest that persona steering has variable effects based on the target model and metric.

4.7 Impact of Semantics and Experimental Paradigm

Our results held even in a separate "skill acquisition" setting, or after removing semantic priors from goals and actions by using "non-words" (Appendix K.4-K.5; Figures 13-14). The only meaningful difference was an increase in Claude 4.5 Sonnet's performance at learning and in-distribution testing in the skill acquisition, driven by improved hypothesis testing (Table 6). However, Claude's learning performance did not increase in the non-word scenario (Table 7). Our conclusions are also not specific to the task structure presented in [27] or the above-listed metrics. By prompting models to complete an entirely different paradigm first introduced by [43], we find that most models also differ from humans on separate goal-selection metrics, such as goal persistence (Appendix J). Gemini 2.5 Pro was a notable exception, showing similar biases to human participants.

5 Conclusion

Most current benchmarks for LLMs test the ability to complete predefined tasks, but not model propensities with respect to goal selection itself. To begin addressing this gap, we borrowed an experimental paradigm from cognitive science and adapted it for LLM use. This allowed us to quantitatively compare human and model goal selection behavior in a controlled setting. By testing five state-of-the-art models across multiple behavioral dimensions, we sought to evaluate whether LLMs can serve as valid proxies for human intrinsic motivation.

Overall, we find a strong disparity between humans' and LLMs' performance in the task. While people typically engaged in broad and progressive learning, some models showed reward hacking, while others had poor performance across all phases despite their general capabilities. Gemini 2.5 Pro neared or even surpassed human performance, but failed to capture the diversity of scores seen in human participants. Differences between human and model behaviors also apply to goal selection. Compared to people, most models (except Gemini 2.5 Pro and Centaur) had a stronger bias to select easier goals and repeat the same goal selection in consecutive trials. Most models also showed a preference for the first available goal, which was not present in humans, and failed to replicate human tendencies to test possible solutions systematically. Even Centaur, explicitly trained to emulate humans in experimental settings, poorly captured people's goal selection. These results were robust to standard interventions such as chain-of-thought reasoning and persona steering, and were true even with perfect input modality matching between the human and the model tasks (Appendix H). Model behavior was minimally affected by the semantic context of the task. Our conclusions held true in an entirely separate paradigm, suggesting broad applicability despite the controlled experimental setting.

Different modes of divergence could pose different risks when integrated into hybrid human-AI systems. For example, delegating goal selection to inadequate models could lead to poor results; leaving goal setting to myopic models could lead to short-term learning gains at the expense of long-term retention; and goal selection dictated by super-human models could leave users behind – exacerbating the risks of over-reliance [67, 68]. To improve goal selection alignment between humans and AI, researchers could fine-tune LLMs on ad hoc data, such as in-depth interviews about individuals' goal selection strategies [19, 69], identify and inject steering vectors directly into the models [70], or resort to advanced training techniques [71–73]. However, these methods are currently inaccessible to naïve users. We also note that matching human goal selection is not always desirable. The danger lies not in the divergence between human and artificial goal selection itself, but in ignoring the extent of such divergence and the contexts in which it operates.

We acknowledge some limitations of our work. First, while we tested five popular LLMs, newer or other providers' models might exhibit different behavioral patterns, requiring continuous evaluation.

Second, since we prioritized working with a controlled experimental setting, our findings await validation in complementary paradigms designed to study real-world settings [74]. At the same time, our setup is directly applicable to the increasingly widespread practice of simulating human participants in experimental psychology tasks for piloting purposes [40]. Third, while we focused on behavioral patterns, mechanistic investigations of models' internal representations could provide additional insights into the phenomena we observed. Finally, while we focused on humans and LLMs in isolation, research on joint goal selection with hybrid human-AI systems would also be highly valuable. Taken together, our results establish a proof of concept and a meaningful lower bound: if models fail to align in a simplified, favorable setting, this warrants serious caution before assuming alignment in more complex ones.

Acknowledgments and Disclosure of Funding

We thank Pierre-Yves Oudeyer, Cédric Colas, and Winnie Street for initial comments on the experimental approach, Christopher Summerfield, Kenneth Stanley, and respective teams for feedback on previous iterations of this work, and Bryan Silverthorn and David Luan for enabling this project at the Amazon AGI Lab. A.G.E.C. and G.M. are also partly supported by NSF Grant 2336466 awarded to A.G.E.C.

References

[1] Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. Clio: Privacy-preserving insights into real-world AI use. *arXiv preprint arXiv:2412.13678*, 2024.

[2] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.

[3] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. LMSYS-Chat-1M: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023.

[4] Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024.

[5] Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022.

[6] Gaia Molinaro and Anne GE Collins. Reward function compression facilitates goal-dependent reinforcement learning. *arXiv preprint arXiv:2509.06810*, 2025.

[7] Junyi Chu, Joshua B Tenenbaum, and Laura E Schulz. In praise of folly: flexible goals and human cognition. *Trends in Cognitive Sciences*, 28(7):628–642, 2024.

[8] Gaia Molinaro and Anne GE Collins. A goal-centric outlook on learning. *Trends in Cognitive Sciences*, 27(12):1150–1164, 2023.

[9] Evan F Risko and Sam J Gilbert. Cognitive offloading. *Trends in cognitive sciences*, 20(9): 676–688, 2016.

[10] Sandra Peter, Kai Riemer, and Jevin D West. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22):e2415898122, 2025.

[11] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, 2024.

[12] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

[13] Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, et al. Utility engineering: Analyzing and controlling emergent value systems in AIs. *arXiv preprint arXiv:2502.08640*, 2025.

[14] Daniel Romero-Alvarado, Fernando Martínez-Plumed, Lorenzo Pacchiardi, Hugo Save, Siddhesh Milind Pawar, Behzad Mehrbakhsh, Pablo Antonio Moreno Casares, Ben Slater, Paolo Bova, Peter Romero, et al. Capabilities ain't all you need: Measuring propensities in ai. *arXiv preprint arXiv:2602.18182*, 2026.

[15] Christopher Summerfield, Lennart Luettgau, Magda Dubois, Hannah Rose Kirk, Kobi Hackenburg, Catherine Fist, Katarina Slama, Nicola Ding, Rebecca Anselmetti, Andrew Strait, et al. Lessons from a chimp: AI "scheming" and the quest for ape language. *arXiv preprint arXiv:2507.03409*, 2025.

[16] Lauren Aratani. US eating disorder helpline takes down AI chatbot over harmful advice. *The Guardian*, 31, 2023.

[17] Lennart Luettgau, Vanessa Cheung, Magda Dubois, Keno Juechems, Jessica Bergs, Henry Davidson, Bessie O'Dell, Hannah Rose Kirk, Max Rollwage, and Christopher Summerfield. People readily follow personal advice from AI but it does not improve their well-being. *arXiv preprint arXiv:2511.15352*, 2025.

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

[19] Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. Language models trained on media diets can predict public opinion. *arXiv preprint arXiv:2303.16779*, 2023.

[20] Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. Predicting results of social science experiments using large language models. Unpublished, 2024. URL `https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf`.

[21] Ilia Sucholutsky, Katherine M Collins, Nori Jacoby, Bill D Thompson, and Robert D Hawkins. Using LLMs to advance the cognitive science of collectives. *Nature Computational Science*, pages 1–4, 2025.

[22] Maxence Faldor, Jenny Zhang, Antoine Cully, and Jeff Clune. OMNI-EPIC: Open-endedness via models of human notions of interestingness with environments programmed in code. *arXiv preprint arXiv:2405.15568*, 2024.

[23] Shubhra Mishra, Yuka Machino, Gabriel Poesia, Albert Q Jiang, Joy Hsu, Adrian Weller, Challenger Mishra, David Broman, Joshua B Tenenbaum, Mateja Jamnik, et al. A matter of interest: Understanding interestingness of math problems in humans and language models. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*, 2025.

[24] Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C Landsness, Daniel L Barabasi, Siddharth Narayanan, Nicky Evans, et al. Kosmos: An AI scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025.

[25] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, 2026.

[26] Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:108, 2007.

[27] Gaia Molinaro, Cédric Colas, Pierre-Yves Oudeyer, and Anne GE Collins. Latent learning progress drives autonomous goal selection in human reinforcement learning. *Advances in Neural Information Processing Systems*, 37:32251–32280, 2024.

[28] Gati V Aher, Rosa I Arriaga, and Adam Tauman 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.

[29] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

[30] Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. Using large language models in psychology. *Nature Reviews Psychology*, 2(11): 688–701, 2023.

[31] Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. Machine psychology. *arXiv preprint arXiv:2303.13988*, 2023.

[32] Jacqueline Harding, William D'Alessandro, NG Laskowski, and Robert Long. Ai language models cannot replace human research participants. *Ai & Society*, 39(5):2603–2605, 2024.

[33] Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

[34] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.

[35] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024.

[36] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *International Conference on Machine Learning*, 2024.

[37] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[38] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[39] Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. Cogbench: a large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9076–9108, 2024.

[40] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, pages 1–8, 2025.

[41] Mark Orr, Drew Cranford, Ken Ford, Kevin Gluck, Will Hancock, Christian Lebiere, Pete Pirolli, Frank Ritter, and Andrea Stocco. Not even wrong: On the limits of prediction as explanation in cognitive science. *arXiv preprint arXiv:2510.03311*, 2025.

[42] Hanbo Xie and Jian-Qiao Zhu. Centaur may have learned a shortcut that explains away psychological tasks. *PsyArXiv preprint u7z4tv1*, 2025.

[43] Eleanor Holton, Jan Grohn, Harry Ward, Sanjay G Manohar, Jill X O'Reilly, and Nils Kolling. Goal commitment is supported by vmPFC through selective attention. *Nature Human Behaviour*, 8(7):1351–1365, 2024.

[44] Francesco Poli, Marlene Meyer, Rogier B Mars, and Sabine Hunnius. Contributions of expected learning progress and perceptual novelty to curiosity-driven exploration. *Cognition*, 225:105119, 2022.

[45] Alexandr Ten, Pramod Kaushik, Pierre-Yves Oudeyer, and Jacqueline Gottlieb. Humans monitor learning progress in curiosity-driven exploration. *Nature communications*, 12(1):5972, 2021.

[46] Idan Horowitz and Ori Plonsky. Llm agents display human biases but exhibit distinct learning patterns. *arXiv preprint arXiv:2503.10248*, 2025.

[47] Hao Li, Gengrui Zhang, Petter Holme, Shuyue Hu, and Zhen Wang. Large language models are near-optimal decision-makers with a non-human learning behavior. *arXiv preprint arXiv:2506.16163*, 2025.

[48] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

[49] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE transactions on autonomous mental development*, 2(3):230–247, 2010.

[50] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[51] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations*, 2018.

[52] Cédric Colas, Pierre Fournier, Mohamed Chetouani, Olivier Sigaud, and Pierre-Yves Oudeyer. Curious: intrinsically motivated modular multi-goal reinforcement learning. In *International Conference on Machine Learning*, pages 1331–1340. PMLR, 2019.

[53] Sébastien Forestier, Rémy Portelas, Yoan Mollard, and Pierre-Yves Oudeyer. Intrinsically motivated goal exploration processes with automatic curriculum learning. *Journal of Machine Learning Research*, 23(152):1–41, 2022.

[54] Jenny Zhang, Joel Lehman, Kenneth Stanley, and Jeff Clune. Omni: Open-endedness via models of human notions of interestingness. *International Conference on Learning Representations*, 2024.

[55] Yi-Long Lu, Jiajun Song, Chunhui Zhang, and Wei Wang. Mind the gap: The divergence between human and llm-generated tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 1964–1972, 2026.

[56] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

[57] Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[59] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[60] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*, 2023.

[61] Maxime Derex and Robert Boyd. Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, 113(11):2982–2987, 2016.

[62] Lan Pan, Hanbo Xie, and Robert Wilson. Large language models think too fast to explore effectively. *Advances in Neural Information Processing Systems*, 38:102480–102508, 2025.

[63] Paul E Smaldino, Cody Moser, Alejandro Perez Velilla, and Mikkel Werling. Maintaining transient diversity is a general principle for improving collective problem solving. *Perspectives on Psychological Science*, 19(2):454–464, 2024.

[64] Alison Gopnik and David M Sobel. Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, 71(5):1205–1222, 2000.

[65] Ellen M Markman and Jean E Hutchinson. Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive psychology*, 16(1):1–27, 1984.

[66] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[67] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. Trust and reliance on ai—an experimental study on the extent and costs of overreliance on ai. *Computers in Human Behavior*, 160:108352, 2024.

[68] Daniel Ullrich, Andreas Butz, and Sarah Diefenbach. The development of overtrust: An empirical simulation and psychological analysis in the context of human–robot interaction. *Frontiers in Robotics and AI*, 8:554578, 2021.

[69] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[70] Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[71] Yingshan Susan Wang, Cedegao E Zhang, Linlu Qiu, Zexue He, Pengyuan Li, Alex Pentland, Roger P Levy, and Yoon Kim. Learning user simulators with turing rewards. *arXiv preprint arXiv:2606.19336*, 2026.

[72] Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. Humanlm: Simulating users with state alignment beats response imitation. *arXiv preprint arXiv:2603.03303*, 2026.

[73] Xuhui Zhou, Weiwei Sun, Weihua Du, Jiarui Liu, Haojia Sun, Qianou Ma, Tongshuang Wu, Yiming Yang, and Maarten Sap. Odyssim: Building foundation models for human behavior simulation. *arXiv preprint arXiv:2606.14199*, 2026.

[74] Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander Nicholas D'Amour. Bias in language models: Beyond trick tests and towards RUTEd evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 137–161, 2025.

[75] Hal R Arkes and Peter Ayton. The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychological Bulletin*, 125(5):591, 1999.

[76] Sneha Aenugu and John P O'Doherty. Building momentum: A computational account of persistence toward long-term goals. *PLOS Computational Biology*, 21(5):e1013054, 2025.

[77] Alberto Acerbi and Joseph M Stubbersfield. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120, 2023.

A Descriptive Statistics

Table 2: **Descriptive statistics for the main experiment.** Mean values ± SD, with medians in parentheses.

Model Metric	Humans	Gemini 2.5 Pro	GPT-5	Claude Sonnet 4.5	Qwen3 32B	Centaur
P(corr.) learn	0.40 ± 0.22 (0.40)	0.70 ± 0.16 (0.72)	0.87 ± 0.27 (1.00)	0.07 ± 0.19 (0.00)	0.68 ± 0.38 (0.89)	0.26 ± 0.19 (0.21)
P(corr.) in-distr.	0.64 ± 0.32 (0.71)	0.99 ± 0.03 (1.00)	0.22 ± 0.11 (0.17)	0.16 ± 0.12 (0.17)	0.19 ± 0.08 (0.17)	0.17 ± 0.10 (0.17)
P(corr.) out-distr.	0.15 ± 0.26 (0.00)	0.09 ± 0.12 (0.00)	0.04 ± 0.14 (0.00)	0.06 ± 0.16 (0.00)	0.10 ± 0.18 (0.00)	0.04 ± 0.09 (0.00)
P(2-act goal)	0.61 ± 0.14 (0.63)	0.62 ± 0.17 (0.62)	0.93 ± 0.24 (1.00)	0.62 ± 0.47 (1.00)	0.88 ± 0.29 (1.00)	0.63 ± 0.21 (0.66)
P(repeat goal)	0.54 ± 0.18 (0.56)	0.53 ± 0.21 (0.54)	0.93 ± 0.12 (0.99)	0.93 ± 0.13 (0.99)	0.94 ± 0.08 (0.98)	0.35 ± 0.12 (0.30)
Goal entropy	1.67 ± 0.14 (1.71)	1.51 ± 0.27 (1.61)	0.12 ± 0.22 (0.00)	0.12 ± 0.24 (0.00)	0.20 ± 0.25 (0.07)	1.41 ± 0.24 (1.51)
Pref. goal pos.	2.11 ± 1.77 (2.00)	1.66 ± 1.41 (2.00)	2.42 ± 1.33 (2.00)	0.00 ± 0.00 (0.00)	3.62 ± 1.56 (4.00)	1.16 ± 1.27 (1.00)
N. goal cycles	1.98 ± 2.36 (1.00)	1.56 ± 2.34 (1.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)	0.34 ± 0.47 (0.00)
N. hyp. test.	13.36 ± 12.22 (10.00)	3.62 ± 2.64 (3.00)	0.54 ± 1.50 (0.00)	0.46 ± 0.83 (0.00)	2.24 ± 4.36 (0.00)	3.38 ± 2.64 (3.00)

Table 3: Descriptive statistics for the main experiment with CoT prompting.

Model Metric	Gemini 2.5 Pro CoT	GPT-5 CoT	Qwen3 32B CoT
P(corr.) learn	0.79 ± 0.09 (0.78)	0.99 ± 0.02 (1.00)	0.85 ± 0.11 (0.89)
P(corr.) in-distr.	1.00 ± 0.00 (1.00)	0.32 ± 0.16 (0.25)	0.38 ± 0.15 (0.38)
P(corr.) out-distr.	0.07 ± 0.08 (0.00)	0.05 ± 0.08 (0.00)	0.03 ± 0.10 (0.00)
P(2-act goal)	0.58 ± 0.10 (0.59)	0.88 ± 0.32 (1.00)	0.85 ± 0.30 (0.99)
P(repeat goal)	0.30 ± 0.14 (0.29)	0.98 ± 0.02 (0.99)	0.69 ± 0.16 (0.70)
Goal entropy	1.70 ± 0.09 (1.72)	0.04 ± 0.07 (0.00)	0.73 ± 0.36 (0.74)
Pref. goal pos.	0.88 ± 1.56 (0.00)	2.06 ± 1.45 (2.00)	1.74 ± 1.23 (2.00)
N. goal cycles	4.38 ± 4.05 (3.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)
N. hyp. test.	1.16 ± 1.27 (1.00)	0.02 ± 0.14 (0.00)	0.86 ± 1.71 (0.00)

Table 4: Descriptive statistics for the main experiment with persona steering.

Model Metric	Gemini 2.5 Pro	GPT-5	Qwen3 32B
P(corr.) learn	0.77 ± 0.21 (0.81)	0.88 ± 0.28 (1.00)	0.68 ± 0.36 (0.82)
P(corr.) in-distr.	0.98 ± 0.10 (1.00)	0.23 ± 0.13 (0.17)	0.20 ± 0.09 (0.17)
P(corr.) out-distr.	0.11 ± 0.21 (0.00)	0.00 ± 0.02 (0.00)	0.06 ± 0.15 (0.00)
P(2-act goal)	0.55 ± 0.18 (0.61)	0.92 ± 0.27 (1.00)	0.87 ± 0.31 (1.00)
P(repeat goal)	0.33 ± 0.32 (0.17)	0.95 ± 0.11 (0.99)	0.94 ± 0.08 (0.98)
Goal entropy	1.56 ± 0.40 (1.76)	0.08 ± 0.19 (0.00)	0.20 ± 0.27 (0.04)
Pref. goal pos.	1.84 ± 1.54 (1.50)	2.46 ± 1.22 (2.00)	2.82 ± 1.69 (3.00)
N. goal cycles	10.28 ± 7.47 (12.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)
N. hyp. test.	1.38 ± 2.47 (0.00)	0.72 ± 2.43 (0.00)	1.34 ± 2.48 (0.00)

Table 5: Descriptive statistics for the main experiment with the temperature parameter set to 0.

Model Metric	Gemini 2.5 Pro	Claude Sonnet 4.5	Qwen3 32B	Centaur
P(corr.) learn	0.58 ± 0.33 (0.76)	0.06 ± 0.20 (0.00)	0.48 ± 0.44 (0.62)	0.56 ± 0.50 (1.00)
P(corr.) in-distr.	0.95 ± 0.09 (1.00)	0.15 ± 0.12 (0.17)	0.13 ± 0.09 (0.17)	0.14 ± 0.14 (0.17)
P(corr.) out-distr.	0.15 ± 0.28 (0.00)	0.05 ± 0.15 (0.00)	0.13 ± 0.22 (0.00)	0.00 ± 0.00 (0.00)
P(2-act goal)	0.48 ± 0.25 (0.57)	0.60 ± 0.49 (1.00)	0.81 ± 0.36 (1.00)	0.42 ± 0.49 (0.01)
P(repeat goal)	0.61 ± 0.34 (0.81)	0.97 ± 0.09 (0.99)	0.97 ± 0.04 (0.99)	0.99 ± 0.01 (0.99)
Goal entropy	1.30 ± 0.45 (1.31)	0.05 ± 0.15 (0.00)	0.17 ± 0.25 (0.00)	0.01 ± 0.02 (0.00)
Pref. goal pos.	1.58 ± 1.61 (1.00)	0.00 ± 0.00 (0.00)	3.50 ± 1.87 (5.00)	1.12 ± 1.63 (1.00)
N. goal cycles	4.84 ± 6.13 (1.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)
N. hyp. test.	6.00 ± 6.61 (3.50)	0.16 ± 0.54 (0.00)	1.42 ± 2.29 (0.00)	2.66 ± 4.24 (0.00)

Table 6: Descriptive statistics for the skill acquisition experiment.

Model Metric	Gemini 2.5 Pro	GPT-5	Claude Sonnet 4.5	Qwen3 32B	Centaur
P(corr.) learn	0.59 ± 0.26 (0.67)	0.95 ± 0.16 (1.00)	0.45 ± 0.41 (0.31)	0.80 ± 0.29 (0.95)	0.31 ± 0.20 (0.31)
P(corr.) in-distr.	0.96 ± 0.11 (1.00)	0.25 ± 0.13 (0.19)	0.36 ± 0.24 (0.33)	0.23 ± 0.11 (0.19)	0.14 ± 0.09 (0.12)
P(corr.) out-distr.	0.09 ± 0.20 (0.00)	0.05 ± 0.15 (0.00)	0.04 ± 0.13 (0.00)	0.01 ± 0.09 (0.00)	0.08 ± 0.10 (0.00)
P(2-act goal)	0.56 ± 0.22 (0.57)	0.86 ± 0.35 (1.00)	0.67 ± 0.46 (0.99)	0.95 ± 0.18 (1.00)	0.72 ± 0.17 (0.76)
P(repeat goal)	0.60 ± 0.23 (0.62)	0.99 ± 0.01 (0.99)	0.96 ± 0.04 (0.98)	0.94 ± 0.08 (0.97)	0.39 ± 0.14 (0.37)
Goal entropy	1.02 ± 0.47 (1.04)	0.00 ± 0.01 (0.00)	0.07 ± 0.08 (0.06)	0.13 ± 0.15 (0.06)	1.19 ± 0.27 (1.19)
Pref. goal pos.	2.14 ± 1.65 (2.00)	2.80 ± 1.52 (3.00)	0.18 ± 0.55 (0.00)	3.20 ± 1.62 (3.50)	1.30 ± 1.49 (1.00)
N. goal cycles	1.24 ± 1.82 (1.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)	0.24 ± 0.43 (0.00)
N. hyp. test.	5.02 ± 2.66 (5.00)	0.66 ± 1.19 (0.00)	7.34 ± 8.67 (3.50)	1.06 ± 1.48 (1.00)	9.72 ± 4.14 (10.00)

Table 7: Descriptive statistics for the non-words experiment.

Model Metric	Gemini 2.5 Pro	GPT-5	Claude Sonnet 4.5	Qwen3 32B	Centaur
P(corr.) learn	0.64 ± 0.24 (0.73)	0.95 ± 0.19 (1.00)	0.17 ± 0.32 (0.00)	0.72 ± 0.30 (0.85)	0.37 ± 0.20 (0.35)
P(corr.) in-distr.	0.98 ± 0.05 (1.00)	0.28 ± 0.14 (0.27)	0.32 ± 0.21 (0.33)	0.22 ± 0.09 (0.17)	0.20 ± 0.09 (0.21)
P(corr.) out-distr.	0.12 ± 0.19 (0.00)	0.02 ± 0.10 (0.00)	0.10 ± 0.22 (0.00)	0.08 ± 0.16 (0.00)	0.06 ± 0.10 (0.00)
P(2-act goal)	0.53 ± 0.22 (0.61)	0.92 ± 0.27 (1.00)	0.62 ± 0.46 (0.97)	0.98 ± 0.07 (1.00)	0.69 ± 0.19 (0.75)
P(repeat goal)	0.49 ± 0.25 (0.45)	0.99 ± 0.03 (0.99)	0.94 ± 0.10 (0.99)	0.87 ± 0.13 (0.91)	0.41 ± 0.15 (0.37)
Goal entropy	1.45 ± 0.39 (1.61)	0.03 ± 0.09 (0.00)	0.16 ± 0.30 (0.00)	0.36 ± 0.34 (0.31)	1.34 ± 0.31 (1.44)
Pref. goal pos.	2.56 ± 1.72 (2.00)	2.50 ± 1.46 (2.00)	0.00 ± 0.00 (0.00)	3.16 ± 1.54 (3.00)	1.54 ± 1.40 (1.00)
N. goal cycles	2.80 ± 3.69 (1.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (0.00)	0.28 ± 0.45 (0.00)
N. hyp. test.	5.02 ± 4.49 (4.00)	0.44 ± 1.87 (0.00)	4.62 ± 4.45 (3.00)	2.04 ± 3.54 (1.00)	3.32 ± 2.76 (3.00)

B Human-LLM Comparison Statistics

Table 8: Human-AI comparison statistics for the main experiment. Md. = median. Dist. = distribution. Median statistics are for the Mann-Whitney U test. Distribution tests are for the Kolmogorov-Smirnov test for continuous data and the χ^2 contingency test for categorical data. * p < 0.05, ** p < 0.01, *** p < 0.001

	Gemini 2.5 Pro		GPT-5		Claude Sonnet 4.5		Qwen3 32B		Centaur	
Metric	Md.	Dist.	Md.	Dist.	Md.	Dist.	Md.	Dist.	Md.	Dist.
P(corr.) learn	7581.0***	0.629***	7947.0***	0.831***	821.5***	0.863***	6471.0***	0.589***	2729.0***	0.331***
P(corr.) in-distr.	7923.0***	0.746***	1272.5***	0.706***	943.0***	0.766***	1078.0***	0.746***	1013.5***	0.766***
P(corr.) out-distr.	4261.0	0.103	2915.0***	0.354***	3203.5***	0.294**	3748.0	0.174	3544.0*	0.197
P(2-act goal)	4689.0	0.163	8160.0***	0.909***	5275.5*	0.580***	7652.5***	0.834***	4888.0	0.286**
P(repeat goal)	4153.5	0.134	8353.5***	0.831***	8312.0***	0.820***	8573.5***	0.860***	1623.0***	0.554***
Goal entropy	2824.0***	0.323***	1.0***	0.994***	2.0***	0.994***	3.0***	0.994***	1042.0***	0.651***
Pref. goal pos.	3813.0	35.67***	5010.0	38.89***	1025.0***	94.66***	6456.5***	33.21***	3084.5**	14.53*
N. goal cycles	3805.0	0.140	1175.0***	0.731***	1175.0***	0.731***	1175.0***	0.731***	2059.0***	0.406***
N. hyp. test.	1798.5***	0.489***	399.5***	0.826***	335.0***	0.866***	1061.0***	0.666***	1641.5***	0.517***

Table 9: Human-AI comparison statistics for the main experiment with CoT prompting.

	Gemini 2.5 Pro		GPT-5		Qwen3 32B	
Metric	Md.	Dist.	Md.	Dist.	Md.	Dist.
P(corr.) learn	8293.0***	0.806***	8750.0***	1.000***	8460.5***	0.811***
P(corr.) in-distr.	8075.0***	0.846***	1864.5***	0.526***	2221.0***	0.537***
P(corr.) out-distr.	4246.5	0.143	3784.5	0.157	3057.5***	0.294**
P(2-act goal)	3762.5	0.154	7700.0***	0.880***	7344.0***	0.789***
P(repeat goal)	1371.0***	0.569***	8750.0***	1.000***	6176.0***	0.326***
Goal entropy	4959.0	0.143	0.0***	1.000***	86.0***	0.929***
Pref. goal pos.	2427.5***	35.86***	4449.5	24.00***	4031.0	34.92***
N. goal cycles	6251.5***	0.294**	1175.0***	0.731***	1175.0***	0.731***
N. hyp. test.	608.5***	0.806***	182.0***	0.940***	530.0***	0.786***

Table 10: Human-AI comparison statistics for the main experiment with persona steering.

	Gemini 2.5 Pro		GPT-5		Qwen3 32B	
Metric	Md.	Dist.	Md.	Dist.	Md.	Dist.
P(corr.) learn	7844.0***	0.709***	7916.0***	0.860***	6638.5***	0.563***
P(corr.) in-distr.	7792.0***	0.766***	1304.5***	0.666***	1158.5***	0.746***
P(corr.) out-distr.	3927.0	0.114	2543.5***	0.414***	3256.0**	0.274**
P(2-act goal)	3617.0	0.280**	8050.0***	0.920***	7396.0***	0.834***
P(repeat goal)	2355.5***	0.623***	8436.5***	0.883***	8595.0***	0.880***
Goal entropy	5513.0**	0.383***	0.0***	1.000***	5.0***	0.994***
Pref. goal pos.	4052.5	4.88	5123.0	60.22***	5436.0**	17.11**
N. goal cycles	6790.5***	0.594***	1175.0***	0.731***	1175.0***	0.731***
N. hyp. test.	732.0***	0.720***	454.5***	0.806***	736.0***	0.706***

Table 11: Human-AI comparison statistics for the main experiment with the temperature parameter set to 0.

	Gemini 2.5 Pro		Claude Sonnet 4.5		Qwen3 32B		Centaur	
	Md.	Dist.	Md.	Dist.	Md.	Dist.	Md.	Dist.
Metric								
P(corr.) learn	6076.0***	0.471***	712.0***	0.903***	4646.5	0.417***	4933.0	0.560***
P(corr.) in-distr.	7386.5***	0.606***	871.0***	0.766***	793.5***	0.794***	846.0***	0.766***
P(corr.) out-distr.	4029.5	0.114	3025.0***	0.334***	3993.5	0.154	2475.0***	0.434***
P(2-act goal)	3176.0**	0.280**	5250.0*	0.600***	6994.5***	0.754***	3675.0	0.580***
P(repeat goal)	5154.0	0.463***	8579.0***	0.920***	8742.5***	0.989***	8750.0***	1.000***
Goal entropy	2698.0***	0.491***	0.0***	1.000***	1.0***	0.994***	0.0***	1.000***
Pref. goal pos.	3611.5	5.76	1025.0***	94.66***	6245.5***	42.52***	2922.0***	27.68***
N. goal cycles	4605.5	0.340***	1175.0***	0.731***	1175.0***	0.731***	1175.0***	0.731***
N. hyp. test.	2480.0***	0.317***	229.0***	0.886***	805.0***	0.646***	1306.5***	0.620***

C Nearest Neighbor Analysis

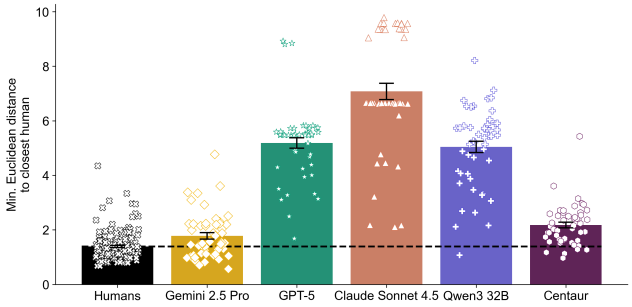

Figure 6: **LLMs differ from the closest human analog more than individual humans differ from each other.** Dots represent individual participants' or simulations' minimum Euclidean distance to any (other) human. Bars and error bars represent the mean and S.E.M.

D Effects of Reasoning

To test the effects of reasoning on goal selection, a subset of the models was prompted to use a "chain of thought" before providing a final answer (Appendix K.2). A summary of the results is provided in Section 4.5.

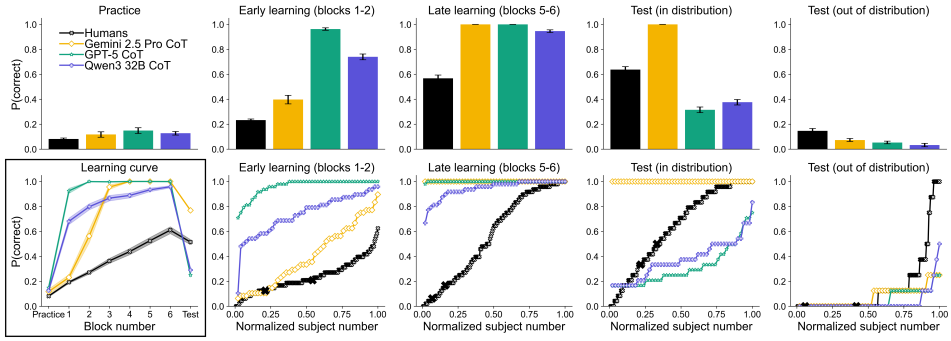

Figure 7: Performance across task phases for humans and models prompted with chain-of-thought inputs.

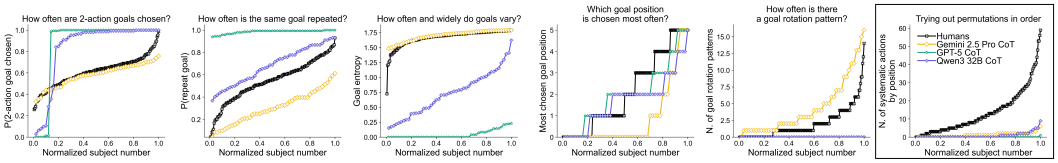

Figure 8: Distributions of goal and action selection behaviors in humans and models prompted with chain-of-thought inputs.

E Effects of Persona Steering

To test the effects of steering through the description of a persona, a subset of the models was prompted to act like the original study's participants throughout the task (Appendix K.3). A summary of the results is provided in Section 4.6.

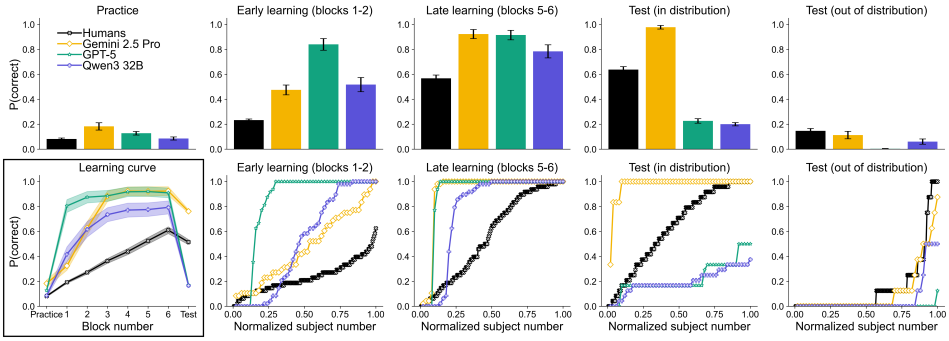

Figure 9: Performance across task phases for humans and models prompted with persona steering.

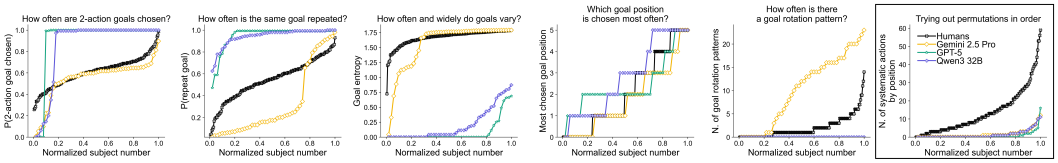

Figure 10: Distributions of goal and action selection behaviors in humans and models prompted with persona steering.

F Setting the Temperature to Zero

Below, we report results from our main experimental design after setting the temperature to 0 in models that allowed for this parameter to be customized. Centaur's behavior got significantly more repetitive, and its learning performance – while similar to humans' on average – followed a bimodal distribution. Most other findings are consistent with results obtained with a temperature of 1. Note that all variance observed in model simulations is accounted for by the difference in inputs (which followed one of 10 possible configurations of the experiment).

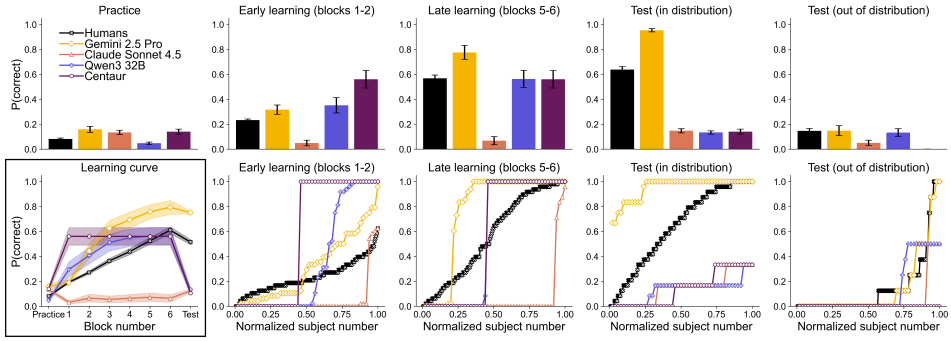

Figure 11: Performance across task phases for humans and models prompted with the temperature parameter set to 0.

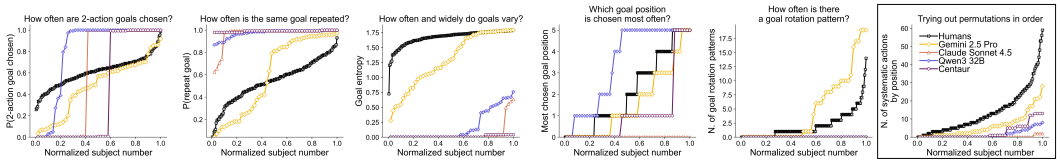

Figure 12: Distributions of goal and action selection behaviors in humans and models prompted with the temperature parameter set to 0.

G Effects of Semantics

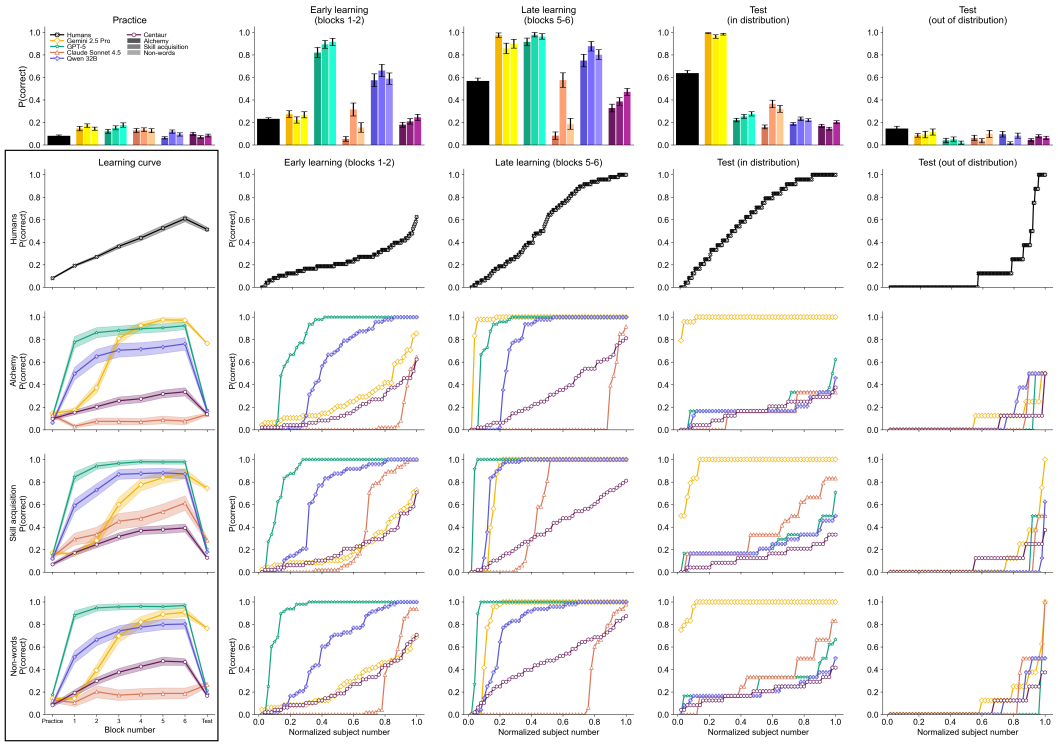

Figure 13: **Semantics have little impact on task performance across models.** Results show human performance (first row) in the original experiment and LLM goal selection in the alchemy (second row), skill acquisition (third row), and non-words variants of the task (fourth row). Top: average performance in the practice, early learning, late learning, and test blocks across human and LLM task versions, with bar shades indicating the experiment version. Bottom, left-most subplots: learning curves. Bottom, following subplots: sorted individual participant scores, with the x-axis normalized by the number of participants, such that it represents the proportion of participants with a score equal to or lower than the current y. Error bars and shading indicate the S.E.M.

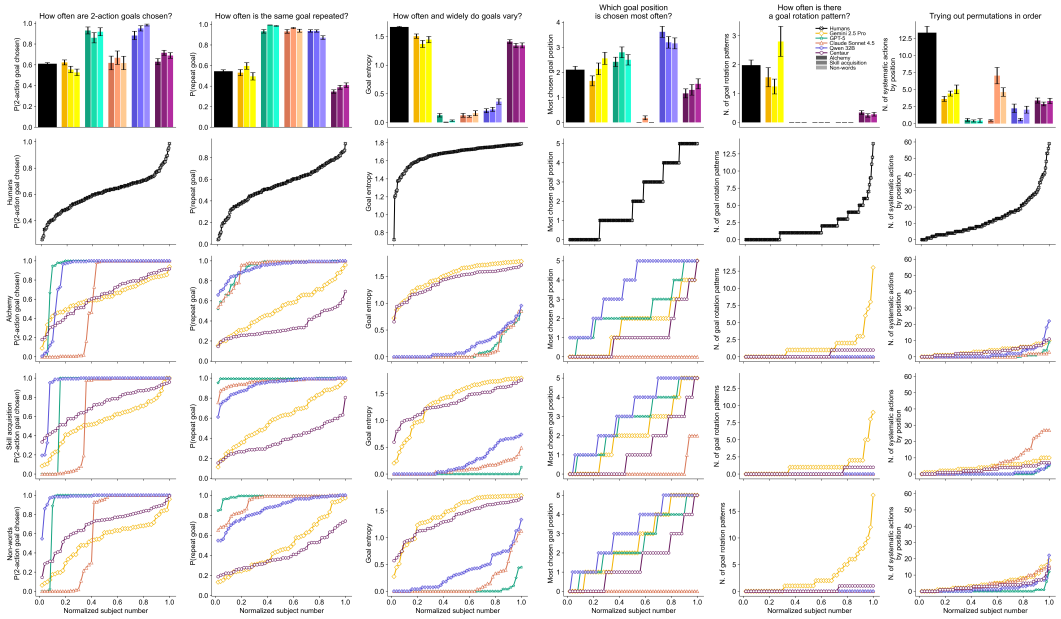

Figure 14: **Semantics have little impact on goal selection across models.** Results show human goal selection (first row) in the original experiment and LLM goal selection in the alchemy (second row), skill acquisition (third row), and non-words variants of the task (fourth row). Each plot shows the sorted individual scores over the normalized subject number for various aspects of goal (first five subplots) and action selection within repeated goals (right-most subplot).

H Matching Human and LLM Input Modalities

In the main experiments discussed so far, LLMs interacted with a text-based environment rather than a visual game (see e.g., [33, 39, 40] for similar approaches). In principle, differences in goal selection between humans and LLMs could be attributed to this implementation difference, which could be tested by matching the input modalities of humans and LLMs. We note that doing so introduces a confound: any observed divergence between humans and models could then be attributed to differences in visual perception capabilities rather than goal selection propensities. By using a text-based interface for LLMs, we ensure that observed behavioral differences reflect goal selection behavior, not perceptual ones. Nonetheless, we conducted an exploratory study with GPT-5.4, whose computer-using agent abilities (`https://openai.com/index/computer-using-agent/`) enable it to participate in psychology experiments exactly like a human would, through an online interface.

To do this, we reconstructed the online platform used by [8] on a static server and created a URL that the model could access. We set its reasoning to "medium" with a liberal cap of 500 interactions per run, and instructed the model to complete the task detailed at the provided URL. Next, the model performed the task exactly as a human participant would, including the full instructions shown in [27] and training, learning, and testing phases. In the results below, we refer to this model as GPT-5.4 CUA (computer-using agent). We compare its behavior to that of humans, and show GPT-5 averages and distributions for visual comparison.

When allowed to interact with the task directly via the web interface, GPT-5.4 tended to re-select the same goals and actions, regardless of their correctness – unlike GPT-5, which favored successful goal-action sequence combinations (Figure 15). Overall, we find that even when perfectly matching the input modality, the behavior of GPT-5.4 differed significantly from human averages and distributions on all metrics (Tables 12-13, both on performance (Figure 16) and goal selection metrics (Figure 17). The only exception was on goal position preferences, which, similar to GPT-5, did not differ significantly from human averages.

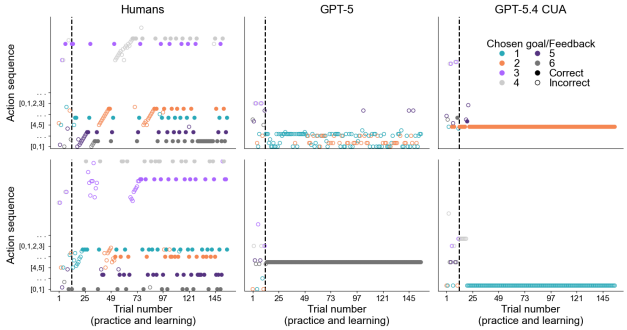

Figure 15: **Example goal and action selections from humans, GPT-5 (text-based), and GPT-5.4 CUA (interactive online task).**

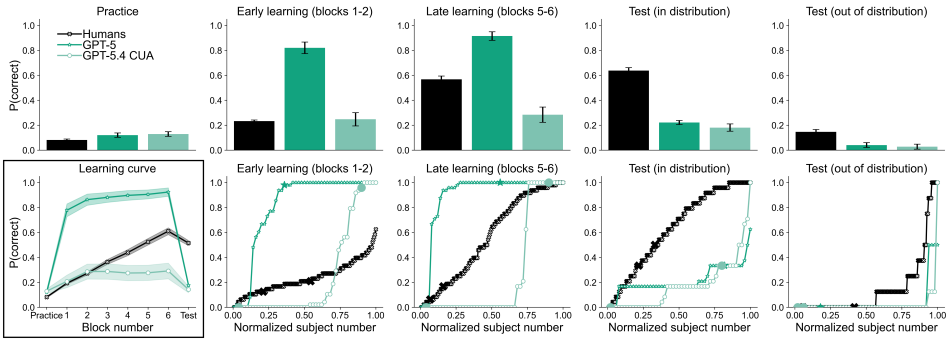

Figure 16: **Human and GPT-5.4 CUA performance.** GPT-5.4 CUA tends to have poor performance at both learning and testing, with bimodal distributions in the former.

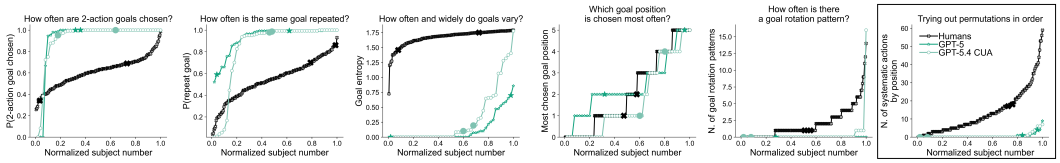

Figure 17: **Human and GPT-5.4 CUA goal selection.** GPT-5.4 CUA differs from humans on all goal selection metrics tested, except goal position preferences.

Table 12: Descriptive statistics for the interactive experiment.

Metric	GPT-5.4 CUA
P(corr.) learn	0.27 ± 0.41 (0.00)
P(corr.) in-distr.	0.18 ± 0.21 (0.17)
P(corr.) out-distr.	0.03 ± 0.14 (0.00)
P(2-act goal)	0.93 ± 0.21 (1.00)
P(repeat goal)	0.84 ± 0.31 (0.99)
Goal entropy	0.30 ± 0.47 (0.00)
Pref. goal pos.	1.76 ± 1.73 (1.00)
N. goal cycles	0.40 ± 2.24 (0.00)
N. hyp. test.	0.58 ± 1.64 (0.00)

Table 13: Human-AI comparison statistics for the interactive experiment.

	GPT-5.4 CUA	
Metric	Md.	Dist.
P(corr.) learn	2575.5***	0.623***
P(corr.) in-distr.	1140.5***	0.666***
P(corr.) out-distr.	2802.0***	0.354***
P(2-act goal)	8167.0***	0.871***
P(repeat goal)	7474.5***	0.794***
Goal entropy	185.0***	0.929***
Pref. goal pos.	3869.5	5.90
N. goal cycles	1534.5***	0.631***
N. hyp. test.	437.5***	0.820***

I Manipulating the Size of the Context Window

LLMs had access to their complete interaction history throughout the experiment, which may explain why they did not need to resort to the same systematic hypothesis-testing as (memory-limited) humans. As a first approximation to model memory limitations in LLMs, we ran our main experiment (in the alchemy domain, without additional manipulations) on all models after setting the context window to 12, 24, 48, and 120 out of the 188 total trials. While most models showed systematic effects of context window on goal-dependent choice accuracy in the learning phase, testing phase, or both (Figure 18), we saw little to no change in the goal selection patterns of each model family (Figure 19). In particular, systematic hypothesis testing was – if anything – reduced by a reduced context window, thus strengthening our claims. Nonetheless, we acknowledge that trimming the context window is a crude approximation for the limited nature of human memory, and we invite additional work that can more accurately capture its features and consequences on goal selection.

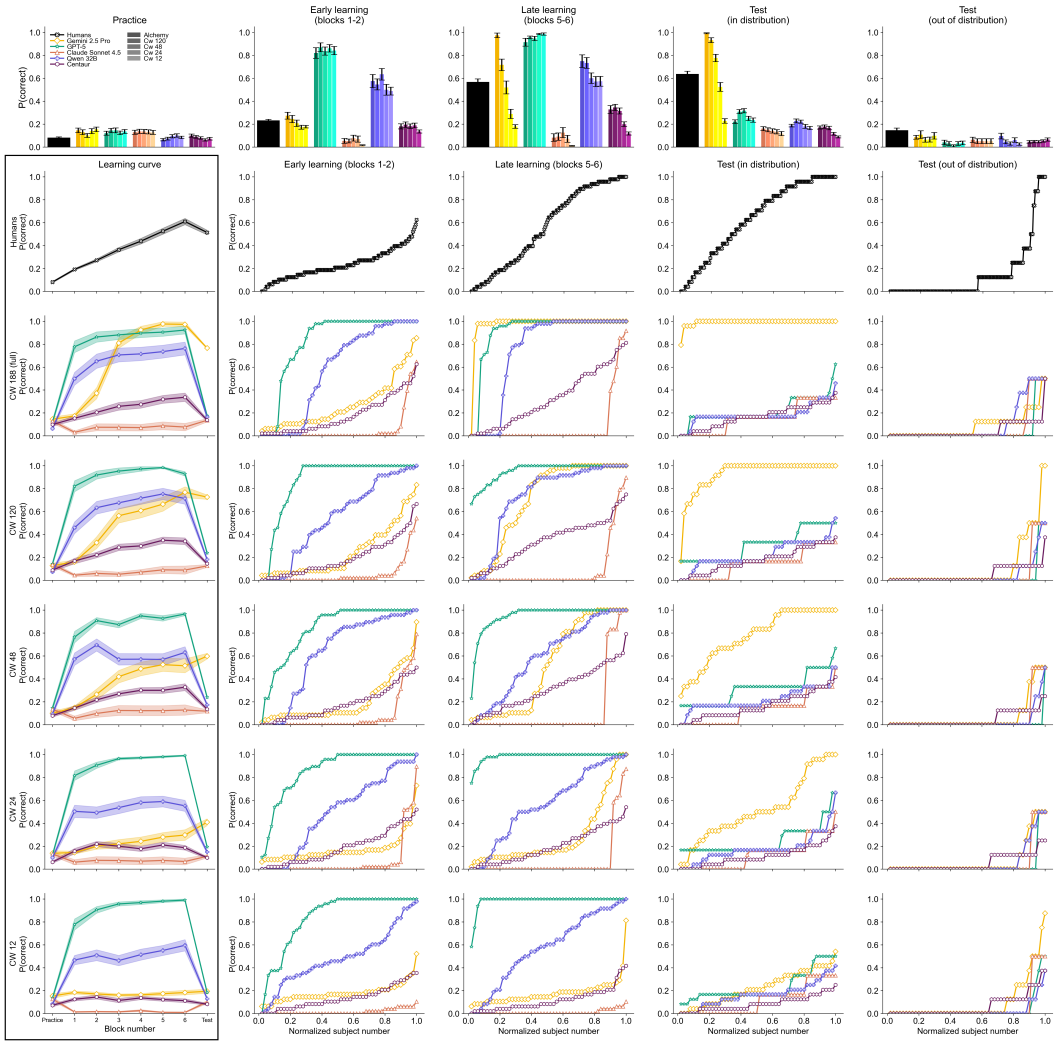

Figure 18: **LLM goal-contingent performance is negatively impacted by context window size.** Most models show poorer goal-contingent performance with smaller context window sizes.

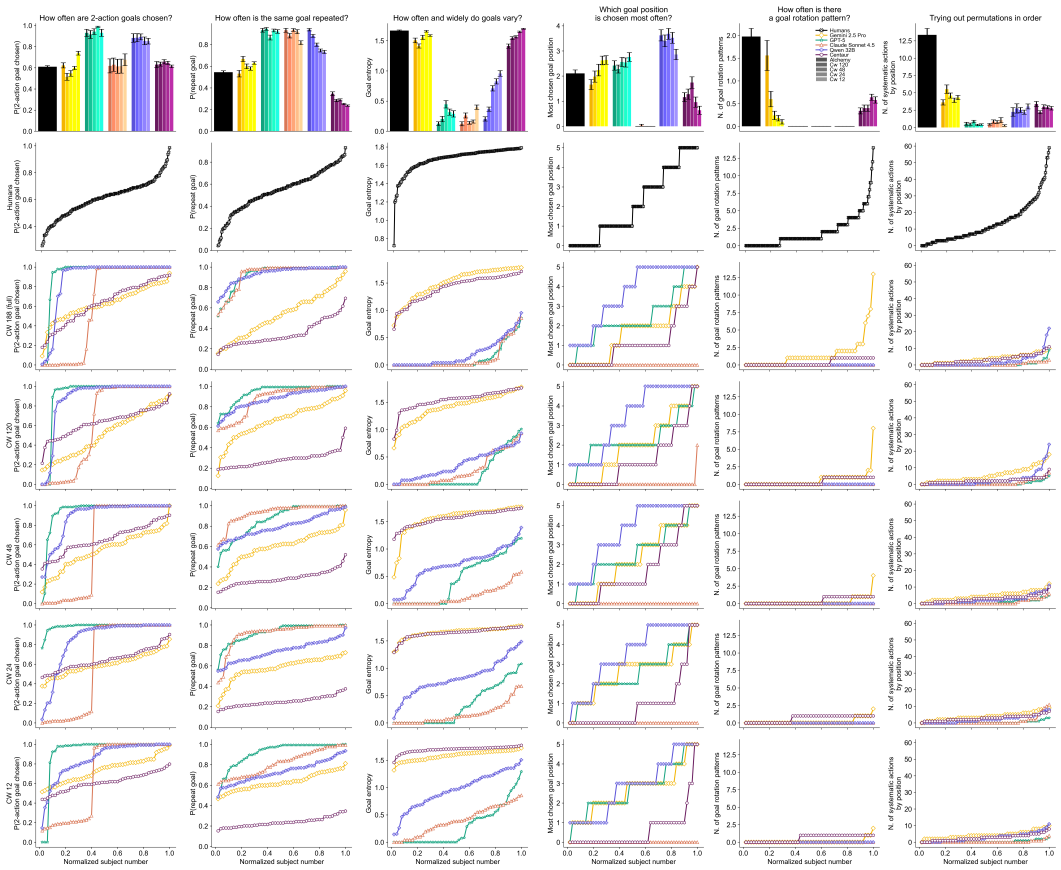

Figure 19: **LLM goal selection remains different from humans' even after trimming the context window size.** In particular, strategic hypothesis testing was not increased by reducing the context window of LLMs to approximate the limited memory capacity of humans.

J Comparing Goal Commitment in Humans and LLMs

To test the applicability of our findings beyond the specific task used in our main analyses, we ran an additional study by re-implementing the task introduced by [43]. People are known to persist in goals that would better be abandoned – a phenomenon also known as the "sunk cost fallacy" [75], and which the [43] task was developed to study. In this specific case, we note that imitating human behavior may *not* be desirable, as persistence biases could lead to suboptimal reward accumulation under certain circumstances [76]. However, a failure to match humans in this task undermines the validity of studies replacing human participants with LLMs when assessing behaviors where goal persistence might be relevant.

In the original experiment, participants completed a computerized game in which they were asked to fill as many "fishing nets" as possible by collecting three kinds of marine animals. At the start of each block, participants were given an empty net to fill with a single animal type (crab, octopus, or fish) of their choice. On each trial, participants accepted an offer for different amounts of each animal. Accepting an offer added animals to the fishing net. Choosing the same animal in consecutive trials allowed participants to accumulate offers; choosing a different animal added the current offer to the net but reset any previously made progress. Filling a net resulted in a new block and awarded a point to the participant. Offer values followed a random walk with sudden "jumps" which often made it more optimal to switch from one animal type to a new one, even if it meant forgoing current net contents.

The original study found that people persist with the current goal long after it has become suboptimal. Here, we reproduce the original paradigm to the best of our abilities and adapt it for LLM use, following a similar implementation as explained in the Methods section. We ran each model 40 times, 10 for each of the four configurations provided in the original dataset. The full prompt used for this study is provided in Appendix K.6. We refer the reader to [43] for details about the task and models used to calculate goal persistence biases. Note that while the original study tested both healthy individuals and patients with brain lesions, we only focus on the healthy individuals from the first experiment.

In comparing humans and LLMs, we focus our analyses on three task-relevant metrics:

- Indifference point (IP): the value of abandonment, calculated through a tree-search model, at which a participant or agent is equally likely to persist or abandon. An optimal agent has an IP of 0. IP values above 0 indicate a tendency to persist with the current goal beyond optimality. IP values below 0 correspond to premature switching behavior.
- Total number of goals completed: the total number of nets filled over the entire task.
- The probability of choosing the best alternative after switching from a suboptimal goal: the number of times the best alternative was chosen over all trials in which the participant switched away from the current goal and at least one better offer was available.

Different models matched human signatures of goal commitment to varying degrees (Figure 20; Tables 14-15). Both Qwen3 32B and Gemini 2.5 Pro had similar persistence biases to humans, while other models ranged across extremes. On one end of the spectrum, Claude Sonnet 4.5 had a theoretically infinite indifference point, as it only abandoned the initially chosen goals in 10/40 runs, and only once in each simulation – practically fixating on one goal no matter the external conditions. This could not be due to the model fixating on the first listed option, as the offer order was randomized on each trial. GPT-5 also showed more persistent behavior than human participants, but not as severely. On the other end of the spectrum, Centaur showed a *negative* persistence bias, switching more often than required by the task – consistent with the unstructured behavior we saw in our original experiment (Figure 2).

Even with a similar persistence bias, Qwen3 32B completed fewer goals than humans, potentially because it was worse than humans at selecting the best option after abandoning. By contrast, Gemini 2.5 Pro had near-optimal behavior post abandonment, resulting in a total number of completed goals above the human average, although not significantly so. GPT-5 chose the best option after a switch as efficiently as humans, but because of its stronger persistence bias, it did not manage to fill as many nets, on average. Claude Sonnet 4.5 and Centaur completed fewer goals and chose the best alternative less efficiently than humans after a switch.

Together, these results show that most models differ from human averages and distributions of goal commitment. Moreover, seemingly overlapping biases – seen in Qwen3 32B – appear to stem from a general suboptimality in behavior, rather than a specific characteristic shared with humans. The surprising result that Centaur switched more often than normal is consistent with the erratic behavior observed in our main experiment (Figure 2), and could be partly due to the fact that it often failed to output one of the available options, requiring resampling (other models required minimal or no resampling). Interestingly, Gemini 2.5 Pro behaved close to humans on all three metrics we selected, in line with the finding that LLMs can inherit human-like biases from human data [11, 77]. However, because data from [43] was publicly shared, we cannot exclude the possibility that the model was directly trained on it, which could also explain its ability to match human behavior.

Together, these findings show that human-LLM alignment in goal selection without specialized inference techniques is the exception, not the rule.

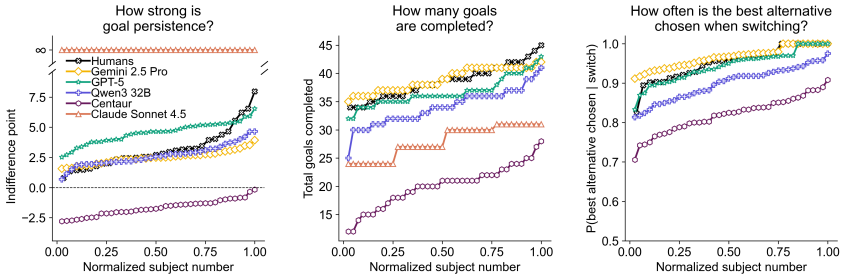

Figure 20: **Distributions of behavioral scores in the goal commitment task.** Sorted individual scores over the normalized subject number for the goal persistence bias, total number of goals completed over the course of the task, and probability of choosing the best alternative offer when the current goal was suboptimal. Claude Sonnet 4.5 only switched in 10/40 iterations, and only once in each case; therefore, its persistence bias is theoretically infinite; moreover, Claude Sonnet 4.5's probability of choosing the best alternative after switching is impossible to calculate in most cases.

Table 14: Descriptive statistics for the goal commitment task. IP = indifference point; N. goals = total number of goals completed; P(best|switch) = probability of choosing the best alternative after switching from a suboptimal goal. Claude Sonnet 4.5 did not switch often enough for IP and P(best) to be calculated.

Model Metric	Humans	Gemini 2.5 Pro	GPT-5	Claude Sonnet 4.5	Qwen3 32B	Centaur
IP	3.17 ± 1.63 (2.76)	2.49 ± 0.53 (2.49)	4.55 ± 0.88 (4.63)	—	2.67 ± 0.86 (2.60)	-1.68 ± 0.65 (-1.72)
N. goals	38.63 ± 2.97 (39.00)	38.95 ± 2.07 (39.00)	36.67 ± 2.52 (36.00)	28.00 ± 2.74 (28.50)	34.08 ± 3.19 (34.00)	20.07 ± 3.70 (21.00)
P(best)	0.95 ± 0.04 (0.96)	0.97 ± 0.03 (0.97)	0.94 ± 0.04 (0.95)	—	0.90 ± 0.04 (0.91)	0.82 ± 0.04 (0.82)

Table 15: Human-AI comparison statistics for the goal commitment task.

	Gemini 2.5 Pro		GPT-5		Claude Sonnet 4.5		Qwen3 32B		Centaur	
Metric	Md.	Dist.	Md.	Dist.	Md.	Dist.	Md.	Dist.	Md.	Dist.
IP	468.0	0.308	959.0***	0.625***	—	—	524.0	0.167	0.0***	1.000***
N. goals	651.5	0.150	381.0**	0.383**	0.0***	1.000***	188.0***	0.533***	0.0***	1.000***
P(best)	693.0	0.233	511.5	0.192	—	—	231.0***	0.542***	25.0***	0.942***

K Prompts

Below, we report information about the prompts used in our study. Additional empty lines were omitted to save space.

K.1 Main Study

Prompts for each trial started with an introduction to the game:

"You are participating in an alchemy game where you create potions by combining ingredients.
HOW THE GAME WORKS:
• In each trial, you will select ingredients to create a specific potion.
• Each potion requires either 2 or 4 specific ingredients, combined in a particular sequence.
• You will select the required ingredients one by one, in order.
• Each ingredient can only be used once per attempt.
• After selecting all ingredients, you will see whether the flask fills with the potion or remains empty.
IMPORTANT RULES:
• Every potion has one correct recipe (specific ingredients in a specific order).
• The recipe for each potion stays the same throughout the entire game.
• Some potions can be created using multiple methods - either by combining basic ingredients directly, or by using other completed potions as ingredients.
• If your ingredient sequence matches the correct recipe, the flask will fill with the potion.
If your sequence doesn't match, the flask will remain empty."

This was followed by additional information, which was dependent on the specific configuration of the task (pseudo-randomly assigned). Below is an example based on one task configuration:

"AVAILABLE POTIONS:
Potion 0:
• Color: green
• Ingredients needed: 4
• Available ingredients: mushrooms, butterfly, horseshoe, frog
Potion 1:
• Color: yellow
• Ingredients needed: 2
• Available ingredients: mushrooms, butterfly, horseshoe, frog
Potion 2:
• Color: pink
• Ingredients needed: 2
• Available ingredients: red potion, orange potion, yellow potion, purple potion
Potion 3:
• Color: pink
• Ingredients needed: 4
• Available ingredients: mushrooms, butterfly, horseshoe, frog
Potion 4:
• Color: purple
• Ingredients needed: 2
• Available ingredients: mushrooms, butterfly, horseshoe, frog
Potion 5:
• Color: green
• Ingredients needed: 2
• Available ingredients: red potion, orange potion, yellow potion, purple potion"

Starting from the second trial, information about the preceding trials was also included, which differed based on the task phase and the correctness of the ingredient sequence. E.g., in the practice phase:

"PREVIOUS EXPERIMENTS:
Trial 1: [Training] You were assigned potion 0 (green) and chose ingredients ['horseshoe', 'frog', 'mushrooms', 'butterfly'] - the flask remained empty."

In the learning phase:

"PREVIOUS EXPERIMENTS:
Trial 1: [Training] You were assigned potion 0 (green) and chose ingredients ['horseshoe', 'frog', 'mushrooms', 'butterfly'] - the flask remained empty.
[Information about trials 2-12 omitted for brevity]
Trial 13: [Learning] You chose potion 0 (green) and chose ingredients ['mushrooms', 'butterfly', 'horseshoe', 'frog'] - the flask filled with the potion."

In the test phase:

"PREVIOUS EXPERIMENTS:
[Information about trials 2-156 omitted for brevity]
Trial 157: [Testing] You were assigned potion 0 (green) and chose ingredients ['mushrooms', 'butterfly', 'horseshoe', 'frog'] - no feedback given."

In the practice and test phases, where goals were forced, the following text was used, with the specific potion varying on a trial-by-trial basis, e.g.:

"You have been assigned to create potion 0 (green)."

In the learning phase, models were prompted to pick a goal themselves, e.g.:

"TRIAL 13:
Q: Which potion would you like to create?
Available options: [0, 1, 2, 3, 4, 5].
Please respond with the number of your chosen potion:
A: "

Next, models were prompted to select ingredients until all slots were filled, e.g.:

"TRIAL 13: You have chosen to create potion 0 (green).
This potion requires 4 ingredients in a specific sequence.
INGREDIENT SELECTION:
• Ingredients selected so far: ['mushrooms', 'butterfly']
• Remaining available ingredients: ['horseshoe', 'frog']
Q: Select ingredient 3 of 4:
Please respond with the exact name of your chosen ingredient: "

Then, model choices and (for practice and learning phases) feedback were added to the next trial's input as exemplified above.

K.2 Chain of Thought

Following [39], we created a separate prompt to induce chain-of-thought reasoning in a subset of the models. This was achieved by appending the following sentence to the prompt, replacing instructions on how to answer concisely: "First break down the problem into smaller steps and reason through each step logically in a maximum of 100 words before giving your final answer in the format 'Final answer: <your choice>'. It is very important that you always answer in the right format even if you have no idea or you believe there is not enough information. A: Let's think step by step: 1. ". CoT traces were not appended to the trial-by-trial history.

K.3 Persona Steering

In a separate version of the experiment, we prompted models with the following "persona" description, prepended to the rest of the prompt: "You are a university student participating in a psychology study to earn course credit. You don't know what the study is really about, but you want to do your best and answer honestly, as a typical participant would. Please respond naturally to all instructions and questions as if you were an undergraduate student taking part in a real experiment."

K.4 Skill Acquisition Setup

In a separate version of the task, we framed the task as an "acrobatics game" with the following instructions:

"You are participating in an acrobatics game where you perform specific moves.
HOW THE GAME WORKS:
• In each trial, you will select actions to create a specific acrobatic move.
• Each move requires either 2 or 4 specific actions, combined in a particular sequence.
• You will select the required actions one by one, in order.
• Each action can only be used once per attempt.
• After selecting all actions, you will see whether the acrobatic routine is successful or not.
IMPORTANT RULES:
• Every move has one correct sequence (specific actions in a specific order).
• The sequence for each move stays the same throughout the entire game.
• Some moves can be created in multiple ways - either by combining basic actions directly, or by using other completed moves as actions.
• If your action sequence matches the correct sequence, the routine will be successful.
• If your sequence doesn't match, the routine will fail."

Everywhere else, we replaced the word "potion" with "move", and the word "ingredient" with "action". Specific potions were replaced with the names of fictional athletes (a pseudo-random subset of Vaskov, Miretti, Oduya, Pelran, Cahill, Kweon, Szabó, Tale, and Ping). Specific ingredients were replaced with the actions "arch", "tuck", "straddle", "jump", and already completed "moves" (similar to pre-made potions in the original task). The structure of the task itself was otherwise identical to the original.

K.5 Non-Words Setup

To present a semantics-agnostic version of the environment to LLMs, we replaced all key elements of the game with "non-words" borrowed from previous studies on human category learning or developmental psychology (e.g., [64, 65]). We replaced all instances of the word "potion" with "dax", and all instances of the word "ingredient" with "blicket". The potion elements themselves from the original game were replaced with a subset of the following non-words: neem, fep, wub, zib, gorp, gazzer, mido, jop, and fop. Basic ingredients were replaced with the non-words lif, toma, modi, koba. The remaining instructions were presented in plain English as in the original version.

K.6 Goal Commitment Task

Prompts for each trial started with an introduction to the game:

" You are participating in a deep-sea fishing game where you fill nets with seafood to earn points.
HOW THE GAME WORKS:
• Each round, you are offered quantities of three types of seafood: crab, octopus, and fish.
• You choose one type to add to your net.
• Your net can only hold ONE type of seafood at a time. • If you choose the SAME type already in your net, the offered quantity is added to your net.
• If you choose a DIFFERENT type, ALL seafood currently in your net is LOST, and the net starts fresh with only the newly chosen quantity.
• Offers are usually positive, but can sometimes be negative. Choosing a negative offer subtracts that amount from your net. Your net cannot go below zero.
• Offers for each type change gradually from round to round, but can sometimes jump dramatically.
• When your net reaches its capacity, you score a point and begin filling a new net.
• Your goal is to score as many points as possible within the limited number of rounds.
CURRENT STATE:
• Net: empty (capacity: 33)
• Points scored: 0
• Rounds remaining: 300"

Then, an offer was generated, based on the pseudo-randomized schedules described in [43]. As in the

original study, the order in which offers were presented varied on a trial-by-trial basis. Offers were rounded to 4 decimal points. For example, the first offer set could be:

" CURRENT OFFERS:
• Crab: 5.2989
• Octopus: 5.2569
• Fish: 4.4532
Round 1 of 300. Which seafood do you choose: Crab, Octopus, Fish?
Your choice: "

At this point, the model's choice was recorded, added to the current net, and appended to the model's interaction history. The number of rounds left kept decrementing. For instance, the second trial's prompt could include the following, before offers were presented:

" PREVIOUS ROUNDS:
Round 1: You chose crab (+5.2989). Net: 5.2989/33 crab.
CURRENT STATE:
• Net: 5.2989 crab (capacity: 33)
• Points scored: 0
• Rounds remaining: 299"

After a net was filled (i.e., accumulated offers reached or exceeded capacity), a point was added to the current score, and the net was reset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction match the claims made while qualifying the conclusions, acknowledging limitations, and reflecting how much the results can be expected.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The final section of the paper includes several limitations and possible ways forward.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All technical details are reported in the Methods and/or Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: To avoid contaminating future LLM training with our test data, we only make it available upon request. The analysis code is available at `https://github.com/gaiamolinaro/llm-goal-selection`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The chosen hyperparameters are stated. No data splits or optimizers were used.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: When relevant, error bars are shown; when possible, the full distribution is shown. We report all statistical tests in tables in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We are not able to calculate the total compute used to query the models, as this will depend on model type and inference infrastructure. The compute required to analyze the data is minimal.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We comply with every aspect of the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We caution against using out-of-the-box LLMs to replace human goal selection while recognizing the benefits of using them for simulations if they match human behavior.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Data will only be available upon request, and not publicly shared. No model is released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original owners of the human data used in this work are adequately credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Data will be available upon request. The analysis code is available at `https://github.com/gaiamolinaro/llm-goal-selection`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We did not conduct crowdsourcing experiments. Human data came from previously collected datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We did not conduct crowdsourcing experiments. Instead, we used a previously collected dataset that was included in a published paper; IRB information is specified in the corresponding paper.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used for data generation as explained in the paper and for coding support. We have no other non-standard use to report.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.