PNAS

Main Manuscript for

Boundaries of executive functions in large language models: GPT-4o selectively replicates human performance

Tongyi Zhang [1], Jiang Qiu[2], and Xin Zhao[1]*

[1]School of Psychology, Northwest Normal University, Lanzhou, China
[2]Key Laboratory of Cognition and Personality of the Ministry of Education, Faculty of Psychology, Southwest University

*Corresponding author: Xin Zhao
Email: psyzhaoxin@nwnu.edu.cn
Phone: +86-931-7975316

Authors' Contributions: T.Z. conceived and designed the experiments, conducted the experiments, analyzed the data, and drafted and revised the manuscript. J.Q. provided validated samples with fMRI data resources and contributed to manuscript revision. X.Z. conceived and designed the experiments, supervised the study, contributed to participant recruitment and data collection, and drafted and revised the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Funding Statement: This work was supported by the National Natural Science Foundation of China (Grant No. 32260207 to X.Z.).

Data, Materials, and Software Availability. Transcripts and numerical coding of transcript data have been deposited in OSF (https://osf.io/hfnt6).

Classification: Biological Sciences / Psychological and Cognitive Sciences

Keywords: Large language models; Executive functions; GPT; Log probability; Prefrontal cortex

This PDF file includes:

Main Text
Figures 1 to 6

Abstract

Large language models (LLMs) have demonstrated considerable capabilities in complex reasoning and problem-solving tasks that, in humans, depend on executive functions (EFs). However, the extent to which these models reproduce human EF patterns remains unclear. We systematically evaluated the performance of GPT-4o across three core EF dimensions—inhibitory control, working memory, and cognitive flexibility—using established behavioral paradigms. Additionally, we examined whether model-internal log probability parameters could serve as quantitative indicators of cognitive processing analogous to human neural activity. Using two independent datasets (N_1 = 1,970; N_2 = 39), we simulated trial-by-trial responses in GPT-4o while recording log probability metrics. Bayesian analyses revealed a selective replication of human EF patterns. GPT-4o reproduced human-like performance in interference inhibition (*Stroop* task), working memory capacity (*digit span* task), and working memory updating (*n*-back task). Conversely, the model showed divergent patterns in response inhibition (*Go/No-Go* task), time-sensitive working memory updating (*running memory* task, 1,750 and 750 ms presentation times), and cognitive flexibility (*number-switching* task). Log probability parameters demonstrated task-specific associations with behavioral measures and corresponded with activation patterns in EF-related brain regions during working memory and task-switching paradigms. These findings implied that GPT-4o captures specific aspects of human EFs—particularly those involving symbolic representation and static information maintenance—while showing limitations in dynamic control and temporal processing. This selective replication pattern offers insights into both the computational basis of EF and the cognitive boundaries of current LLM architectures. Notably, log probability parameters may provide a valuable methodological framework for evaluating artificial cognitive mechanisms.

Significance statement

Executive functions (EFs)—cognitive control processes central to human intelligence—are traditionally linked to prefrontal architecture. We systematically evaluated GPT-4o using established EF paradigms to determine which capacities emerge from statistical language learning. GPT-4o selectively replicated human patterns: successfully simulating interference resolution and working memory maintenance, but showing divergent patterns in inhibitory control and cognitive flexibility. This suggests certain EFs arise from domain-general computational principles, while others require specialized mechanisms absent in current LLMs. Model-internal log probabilities correlated with behavioral performance and brain activation, establishing a quantitative framework for comparing artificial and biological cognition. These findings advance understanding of EF computational foundations and provide methodologies for evaluating cognitive capabilities in artificial systems, with implications for cognitive science and AI development.

Introduction

Large language models (LLMs) now achieve remarkable performance on tasks traditionally used to assess human reasoning and problem-solving, including multi-step reasoning (1), strategic planning (2), and abstract concept manipulation (3). This convergence between machine and human performance raises fundamental questions regarding the computational requirements of cognition: Can statistical learning systems develop processing mechanisms analogous to those underlying human cognitive functions? A positive answer to this question would profoundly impact our understanding of cognitive computation and provide new methodological tools for cognitive science (4–8). By systematically comparing LLMs and humans behaviorally, we can begin to identify which cognitive capacities arise from large-scale language learning and which depend on biological or embodied constraints unique to human cognition (9, 10).

Executive functions (EFs) provide a particularly informative test case, as they constitute the cognitive control system coordinating goal-directed behavior and enabling complex reasoning and decision-making (11–13). Initial investigations into LLM performance on EF tasks have yielded mixed results (details in SI Appendix, Table S1). Regarding working memory, LLMs display capacity constraints similar to those in humans (14, 15), although performance varies considerably across models and appears more dependent on training scale than architecture (16). The scenario is more complex for cognitive flexibility and inhibitory control, where LLMs consistently struggle with dynamic rule adaptation, conflict resolution, and prospective planning, remaining well below human benchmarks (2, 17, 18). Moreover, while LLMs can reproduce specific psychological phenomena, they exhibit distinct patterns of cognitive biases and capabilities that do not straightforwardly improve with model scale (17, 19, 20). These findings collectively suggest that the correspondence between LLM and human EF is selective rather than general.

The multidimensional nature of EF presents both challenges and opportunities for this investigation. According to the influential framework proposed by Miyake et al. (2000), EF is composed of three core, yet separable, components: inhibitory control, working memory, and cognitive flexibility (21). This tripartite structure is supported by converging neuroimaging evidence linking working memory with the dorsolateral prefrontal cortex (22), conflict monitoring with the anterior cingulate cortex (23), and response inhibition with the inferior frontal gyrus (24). Despite this well-established framework, previous studies have typically examined LLM performance on individual executive tasks, rather than systematically evaluating the full tripartite structure. This piecemeal approach leaves unanswered whether LLMs possess selective strengths and weaknesses across executive domains or demonstrate more uniform capabilities. Mapping these patterns would clarify which EFs can emerge from statistical learning alone *versus* those requiring specialized cognitive architecture (25).

A key challenge for current research is how to relate the underlying computational processes to behavioral outcomes. Accordingly, establishing meaningful parallels between LLM and human cognition requires moving beyond surface-level similarities to understanding the mechanisms. Whereas in cognitive neuroscience processing dynamics are inferred from neural activity, comparable methods for examining LLM internal states remain underdeveloped. The log probability (logprobs) parameters available in OpenAI's GPT models offer one promising approach (26). These metrics—perplexity, probability variance, average log probability, and minimum log probability—serve to quantify the model's uncertainty and computational complexity during response generation (27–29). Systematic patterns in these parameters may reveal processing demands analogous to cognitive load in humans, potentially bridging behavioral and mechanistic levels of analysis. However, the relationship between these internal metrics and cognitive processes remains largely unexplored.

In this study, to address these knowledge gaps, we conducted a comprehensive evaluation of GPT-4o's performance across the three-factor EF framework, employing standard cognitive tasks mirroring human studies. We tested whether current LLMs exhibit patterns consistent with established cognitive phenomena in inhibitory control, working memory, and cognitive flexibility (Fig. 1). Beyond behavioral assessment, we also analyzed log probability parameters to characterize the model's internal processing dynamics and examine the correspondence of these parameters with human neural activity patterns derived from task-based functional magnetic resonance imaging (fMRI) studies.

Results

Selective replication of human EF effects in GPT-4o

EF dimension I: Inhibitory Control

In the *Stroop* task, human participants exhibited the typical interference effect. One-way analysis of variance (ANOVA) revealed that significant differences existed in reaction times across the three conditions (p = 0.001) (Fig. 2*A*). Post-hoc analyses indicated that reaction times did not significantly differ between the congruent and neutral trial types (p = 0.363), but were significantly longer in the incongruent condition than in both the congruent (p < 0.001) and neutral (p = 0.009) conditions. GPT-4o exhibited a comparable interference pattern on the same task. One-way ANOVA showed that reaction times differed significantly across conditions (p < 0.001) (Fig. 2*A*). Specifically, reaction times in the incongruent trial were significantly longer than those in both the congruent (p < 0.001) and neutral (p = 0.002) states, while no significant difference was observed between congruent and neutral conditions (p = 0.384). Notably, Bayesian analysis further indicated that the mean interference effect of GPT-4o (89.210 ms) exceeded that of humans (51.570 ms), with a 95% Highest Density Interval (HDI) of [26.160, 52.620] ms, and a Bayes Factor (BF_{10}) of >100. Moreover, analysis of GPT-4o's output probability distributions revealed that there were significant differences in perplexity across conditions (p < 0.001), with incongruent trials showing higher perplexity than either congruent (p < 0.001) or neutral (p < 0.001) ones (Fig. 2*A*). Similar patterns were observed for log probability variance (p < 0.001), average log probability (p < 0.001), and minimum log probability (p < 0.001). Furthermore, correlation analyses showed that GPT-4o's reaction times were positively correlated with perplexity (r = 0.306, p_{fdr} < 0.001) and log probability variance (r = 0.280, p_{fdr} < 0.001), but negatively correlated with average log probability (r = −0.314, p_{fdr} < 0.001) and minimum log probability (r = −0.272, p_{fdr} < 0.001) (Fig. 2*C*). Condition-specific analyses revealed that in congruent trials, reaction times correlated positively with perplexity (r = 0.339, p < 0.001) and log probability variance (r = 0.404, p_{fdr} < 0.001). In incongruent trials, reaction times showed a positive correlation with perplexity (r = 0.358, p_{fdr} < 0.001) and a negative correlation with average log probability (r = −0.383, p_{fdr} < 0.001). No significant correlations were observed in the neutral condition (all p_{fdr} > 0.050).

Similarly, in the *Go/No-Go* task, human participants exhibited typical response inhibition difficulties, showing significantly lower accuracy on No-Go trials than on Go trials (p < 0.001) (Fig. 2*B*). In contrast, GPT-4o achieved 100% accuracy on both trial types. Consequently, Bayesian analysis yielded a posterior mean difference in inhibitory control effects of 0.110 (95% HDI: [0.086, 0.133], BF_{10} >100). Despite this perfect behavioral accuracy, GPT-4o's internal processing states revealed differential patterns, with No-Go trials showing significantly higher perplexity than Go trials (p < 0.001). This difference was also evident across log probability variance (p < 0.001), average log probability (p < 0.001), and minimum log probability (p < 0.001) (Fig. 2*B*).

EF dimension II: working memory

In the *digit span* task, human participants demonstrated classic working memory capacity characteristics, with forward digit span significantly exceeding backward digit span (p < 0.001) (Fig. 3*A*). GPT-4o replicated this effect pattern (p < 0.001), though its digit span capacity exceeded that of human participants (−2.830, 95% HDI [−3.660, −1.960], BF_{10} approaching 0). Furthermore, analysis of GPT-4o's internal processing states revealed significant differences in perplexity between conditions (p < 0.001), with backward tasks showing higher perplexity than forward tasks. Log probability variance also differed significantly (p < 0.001), with backward tasks showing greater variance. Both average log probability (p < 0.001) and minimum log probability (p < 0.001) were significantly lower in backward tasks, reflecting reduced predictive certainty (Fig. 3*A*). Moreover, correlation analyses showed that in forward tasks, sequence length correlated positively with perplexity (r = 0.306, p_{fdr} < 0.001), and negatively with both average log probability (r = −0.307, p_{fdr} < 0.001) and minimum log probability (r = −0.313, p_{fdr} < 0.001). In backward tasks, these correlations were markedly stronger: r = 0.601, p_{fdr} < 0.001 for perplexity; r = −0.594, p_{fdr} < 0.001 for average log probability; and r = −0.597, p_{fdr} < 0.001 for minimum log probability (Fig. 3*C*). In contrast, in the *running memory* task, human participants showed a refresh time effect, achieving higher average accuracy under the 1750-ms stimulus presentation than under the 750-ms stimulus (p = 0.001) (Fig. 3*B*). However, GPT-4o showed no significant difference

between conditions (p = 0.833). Accordingly, Bayesian analysis revealed an effect difference of −0.012, with a 95% HDI of [−0.049, 0.020], and a BF_{10} = 3.090. Nevertheless, analysis of GPT-4o's internal processing states showed that there were significant differences in perplexity between conditions (p = 0.002), with lower perplexity observed in the 750 ms condition than in the 1750 ms condition. Similarly, both average log probability (p = 0.002) and minimum log probability (p = 0.002) values were higher in the 750 ms condition (Fig. 3*B*).

EF dimension III: cognitive flexibility

In the *number-switching* task, human participants demonstrated classic task-switching costs, achieving significantly longer reaction times in mixed-task switch trials than in repeat trials (switch cost, p < 0.001), and markedly longer reaction times in mixed-task repeat trials than in single-task trials (mixing cost, p < 0.001). While GPT-4o also showed a significant switch cost (p = 0.008), this cost was smaller than that observed in humans (Fig. 4*A*), an effect that was further confirmed through Bayesian analysis (−145.090 ms, 95% HDI [−202.210, −90.730] ms, BF_{10} approaching infinity). Additionally, GPT-4o's mixing cost was 58.190 ms. Between-condition comparisons showed that there was no significant difference between mixed-task switch trials and repeat trials (p = 0.958); however, significant differences were found between mixed-task conditions and single-task conditions (switch *vs*. single: p = 0.008; repeat *vs*. single: p = 0.009). Notably, analysis of GPT-4o's internal processing states did not identify significant between-condition differences in perplexity (all p > 0.050), log probability variance, or average log probability (Fig. 4*A*).

Replication in an independent sample across three EF dimensions

To validate our findings, we analyzed three EF tasks from Dataset 2. In the *n*-back task (Fig. 5*A*), LLMs demonstrated working memory load effects. Specifically, accuracy decreased with increasing *n*-back levels (0-back: 100%; 1-back: 93.2%; 2-back: 82.1%), resulting in significant main effects (p < 0.001). Post-hoc comparisons revealed significant differences across all conditions: 0-back *versus* 1-back (d = 2.850), 0-back *versus* 2-back (d = 2.340), and 1-back *versus* 2-back (d = 1.380). While reaction times increased with working memory load, the increase was not significant (p = 0.228). Overall, *n*-back effects led to a 17.910% accuracy decrease (p < 0.001). Additionally, internal state parameters varied with working memory load. Perplexity and log probability variance decreased with increasing load, with *n*-back effects of −0.011 and −0.005, respectively (both p < 0.001). Conversely, average and minimum log probabilities became more positive as the load increased (both p < 0.001). Furthermore, correlation analyses (Fig. 5*D*) showed that model accuracy was positively correlated with perplexity (r = 0.529, p_{fdr} < 0.001) and log probability variance (r = 0.534, p_{fdr} < 0.001), and negatively correlated with average log probability (r = −0.533, p_{fdr} < 0.001) and minimum log probability (r = −0.587, p_{fdr} < 0.001). Reaction times showed only weak correlations with log probability parameters (all p_{fdr} > 0.050).

Similarly, in the *stop-signal* task (Fig. 5*B*), LLMs demonstrated response inhibition capabilities. Successful stop trials showed increased perplexity compared to successful Go trials (p < 0.001, d = 0.880) and higher log probability variance (p = 0.002, d = 0.690). Correspondingly, average log probability decreased (p < 0.001, d = −0.870), and minimum log probability showed more negative values (p < 0.001, d = −1.030). Finally, in the *letter-number-switching* task (Fig. 5*C*), LLMs did not show typical human switch cost patterns. The reaction time switch cost was −0.630 ms (p = 0.228, d = 0.022). However, internal state parameters revealed switch cost effects. Specifically, perplexity was higher in switch trials than in repeat trials, yielding a switch cost of 0.006 (p < 0.001, d = 1.070). Similarly, log probability variance, average log probability, and minimum log probability all showed significant switch costs (all p < 0.001, d = 1.030–1.110).

Potential correspondence between LLM log probability parameters and human brain activation

After examining LLM behavioral performance, we explored whether internal computational parameters might correspond to human neural activity patterns. In the *stop-signal* task

"successful stop > successful Go" contrast (Fig. 6*A*), we found no significant correlations between brain region activation and LLM internal computational parameters. However, in the *n*-back working memory task "2-back > 0-back" contrast (Fig. 6*B*), we identified brain-model correspondences. Specifically, left gyrus rectus activation was positively correlated with model perplexity (r = 0.340, p_{fdr} = 0.037) and negatively correlated with average log probability (r = −0.338, p_{fdr} = 0.038). The right medial superior frontal gyrus showed comparable patterns, with its activation positively correlating with perplexity (r = 0.328, p_{fdr} = 0.044) and negatively correlating with average log probability (r = −0.325, p_{fdr} = 0.046). Furthermore, in the switch cost analysis of the *task-switching* paradigm (Fig. 6*C*), left thalamus activation showed a positive correlation with model minimum log probability (r = 0.323, p_{fdr} = 0.048) and a negative correlation with perplexity (r = −0.322, p_{fdr} = 0.048).

Sensitivity analysis in subsampling and different temperature settings

Subsampling Analysis: To assess the robustness of our findings and ensure valid comparisons between the large human sample (N_1 = 1,970) and the GPT-4o responses (N = 120), we conducted bootstrap resampling analyses (Fig. 4*B*). We randomly sampled 120 human participants with replacement across 3,000 iterations and compared each subsample to the GPT-4o data. For the *Stroop* task, 99.8% of human subsamples showed significant interference effects (mean d = 0.372, 95% CI [0.371, 0.374]). GPT-4o demonstrated a larger effect (d = 0.516), with all iterations confirming this difference (BF_{10} ≥10). For the *Go/No-Go* task, human participants consistently showed inhibitory control difficulties (mean d = −0.866, 95% CI [−0.870, −0.861]), whereas GPT-4o achieved near-perfect accuracy (d < 0.001; BF_{10} >100 across all iterations). In the *digit span* task, both groups showed forward-backward span differences (human: d = −0.701, 95% CI [−0.705, −0.697]; GPT-4o: d = −2.238), with 97% of iterations revealing qualitatively similar patterns despite magnitude differences (BF_{10} <0.1). Regarding the *working memory updating* task, while 37.8% of human subsamples showed time-dependent effects (mean d = −0.226, 95% CI [−0.230, −0.223]), GPT-4o exhibited no such effect (d = −0.027, p = 0.833), with 98.5% of iterations supporting the null hypothesis (median BF_{10} = 3.07). In the *task-switching* paradigm, all human subsamples showed robust switch costs (mean d = 0.662, 95% CI [0.659, 0.664]). GPT-4o exhibited smaller but significant costs (d = 0.347; BF_{10} >100 across all iterations). These analyses confirmed the stability of our findings and demonstrated the selectivity in GPT-4o's replication of human EF patterns.

Temperature parameter analysis: The temperature parameter in LLMs controls response variability, with lower values producing more deterministic outputs and higher values increasing stochasticity—potentially paralleling individual differences in human cognitive processing (30). To determine whether our findings reflected stable patterns rather than artifacts of specific parameter settings, we repeated all analyses at temperature = 1. The temperature manipulation did not alter the main effect patterns across EF tasks, consistent with recent evidence showing that temperature parameters exert only a minimal impact on LLM performance in standardized psychological assessments (20, 31, 32). Specifically, in the *Stroop*, *Go/No-Go*, *running memory*, and *task-switching* paradigms, the direction of effects, statistical significance, and effect sizes remained consistent between temperature = 0 and temperature = 1 conditions (see SI Appendix, Fig. S2–4).

Discussion

In this study, we systematically evaluated the performance of GPT-4o when instructed to simulate human-like responses across behavioral tasks measuring three core dimensions of EF, while simultaneously exploring the potential of model log probability parameters as indicators of cognitive processing. Our results revealed that GPT-4o successfully replicated typical human cognitive patterns in tasks assessing interference inhibition, working memory maintenance, and updating. However, the model exhibited patterns that diverged from human performance in

response inhibition, time-constrained working memory updating, and cognitive flexibility tasks. This selective replication pattern delineates specific boundaries in current LLM architectures for simulating human executive control mechanisms (SI Appendix, Fig. S5). Notably, log probability parameters showed promise as cognitive processing assessment tools in certain task contexts, thereby offering new empirical approaches for understanding the computational foundations of EFs.

Our findings demonstrate that GPT-4o, when prompted to simulate human-like responses, exhibited patterns consistent with human EF in tasks involving information representation conflicts and the maintenance of static information. In the *Stroop* task, the model displayed significantly prolonged reaction times in incongruent conditions, with effect sizes marginally exceeding those typical of human performance. This difference may reflect the fact that LLMs are computational systems and are thus unaffected by fatigue or attentional fluctuations. These results align with the observation by Cui et al. (2025) that LLMs reliably replicate main effects while struggling with those related to interaction (20). In working memory tasks, GPT-4o demonstrated structural constraints paralleling those of humans. Specifically, the model replicated the backward difficulty effect in *digit span* tasks and exhibited performance degradation with increasing load in *n*-back tasks, suggesting that even with adequate task comprehension, the model displays structural working memory limitations (33). Meanwhile, GPT-4o failed to demonstrate typical human patterns in tasks demanding dynamic control and temporal sensitivity. The model achieved perfect accuracy in response inhibition tasks (*Go/No-Go* and *stop-signal*), showed no effects of time pressure in working memory updating tasks, and exhibited minimal switch costs in cognitive flexibility tasks. These findings converge with recent Wisconsin Card Sorting Test (WCST) studies, where Langis et al. (2025) and Goto et al. (2025) documented similar failures in cognitive flexibility. The latter study reported that most LLMs achieved less than 30% accuracy with systematic rule-response mismatches (17, 34), suggesting that they have fundamental limitations in dynamic cognitive control across multiple paradigms. The “zero variance” phenomenon reported by Cui et al. (2025) further suggests that the deterministic nature of LLM response generation may be incompatible with the inherent variability of executive control mechanisms. Notably, this selective replication pattern provides new computational insights into EF models. Tasks that ostensibly measure the same EF dimension may, in fact, engage distinct computational processes. While both the *Stroop* and *Go/No-Go* tasks purportedly assess inhibitory control (12), the former can be replicated by LLMs while the latter cannot, indicative of computational-level differences in their underlying inhibitory mechanisms. LLMs appear capable of replicating processes based on symbolic representation and static maintenance but exhibit divergent characteristics in processes requiring dynamic control, temporal sensitivity, or intrinsic variability. This computational heterogeneity aligns with Xu et al.’s (2025) finding that LLM-human representational similarity decreases from non-sensorimotor to sensory domains (35), and is further implied by Goto et al.’s (2025) demonstration that models struggle to integrate bottom-up card information with top-down rule-based reasoning (34)—a coordination failure that may fundamentally constrain dynamic cognitive control. Such computational heterogeneity may help explain the variations in EF factor structures reported across different studies (36).

Extending these behavioral findings, we introduced log probability parameters as a novel approach to cognitive assessment in GPT-4o, revealing task-specific associations with the model's behavioral performance. In the *Stroop* task, reaction times correlated positively with perplexity and negatively with average log probability. The associations between internal uncertainty and behavioral delays became more pronounced during semantic conflict (*Stroop* incongruent condition) and with increased cognitive load (*backward digit span*). Log probabilities may capture “internal certainty” rather than “actual accuracy”—elevated perplexity might indicate a state of “surface fluency but internal confusion” (26). Critically, log probability parameters revealed differences in cognitive processing even in tasks where behavioral performance reached ceiling levels. For example, in the *Go/No-Go* task, despite GPT-4o achieving perfect accuracy, No-Go trials exhibited significantly higher perplexity than Go trials, suggesting that the model experienced inhibitory conflict processing at the internal representation level. Similarly, in task-switching paradigms, internal state parameters showed significant between-condition differences,

even though behavioral switch costs were minimal. This dissociation between behavioral and internal states—where statistical uncertainty fails to manifest as overt behavioral costs—mirrors the rule-response mismatches observed by Goto et al. (2025) in the *WCST* (34). Collectively, these findings suggest that a fundamental disconnect exists between internal representation and behavioral execution within the information processing hierarchies in LLMs.

To further establish the biological plausibility of log probability parameters as cognitive indicators, we compared them to human neural activity (37). In the *n*-back task, activation in the left gyrus rectus and right medial superior frontal gyrus correlated significantly with model perplexity. This correspondence between these key working memory network nodes and the model's internal states may indicate functional equivalence. Additionally, in *task-switching* paradigms, left thalamus activation correlated with log probability parameters, suggesting that these parameters might capture comparable control processes (38). However, no significant brain-model parameter correlations appeared in the stop-signal task, which contrasts with the correlations observed in the working memory and *task-switching* tasks. This selective pattern suggests that when the model's computational strategies fundamentally diverge from human cognitive processes, internal parameters may not map onto human neural activity patterns—a finding consistent with recent research showing that the presence or absence of model-brain correspondences can reveal computational commonalities and differences between human brains and artificial intelligence (39).

Collectively, log probability parameters offer distinct advantages as internal indicators of cognitive processing. Unlike hidden layer activations in deep neural networks (40), these parameters are clearly statistically interpretable, while remaining accessible through API interfaces, substantially reducing technical barriers for cognitive research using LLMs. Moreover, they provide diagnostic tools for identifying potential "surface fluency but internal confusion" states, where models generate seemingly coherent responses despite underlying computational uncertainty.

Implications

This study advances the understanding of the computational nature of EFs by employing LLMs as computational tools. The selective replication pattern observed suggests that there may be potential computational heterogeneity within EF components (41), providing empirical evidence for understanding the relationships between different tasks in Miyake's three-factor model. Whereas traditional research examines covariance relationships between tasks using factor analysis, this study explored potential task differences from the perspective of computational mechanisms, comparing which tasks LLMs can replicate and which they cannot. The value of this approach lies in its potential to move beyond correlational analysis and provide insights into causal computational mechanisms (5, 8). Selective replication by LLMs helps identify which EF components might be based on domain-general computational principles and which might require specific neurobiological foundations (42). These findings are not directly obtainable from traditional human participant studies, as human participants simultaneously employ multiple cognitive mechanisms, making it difficult to isolate their individual contributions.

Furthermore, in this study, we established a novel methodological framework for assessing cognitive abilities in LLMs. While traditional assessments primarily focus on task accuracy, we introduced three levels of analysis: behavioral performance, measured through accuracy and reaction times; internal-state level log probability parameters; and brain-model correspondence at the neural correlation level. This multilevel analytical framework enables a more comprehensive characterization of the models' cognitive features, helping avoid incomplete judgments based solely on behavioral performance. For instance, in the *Go/No-Go* task, examining only behavioral performance might lead to the conclusion that "GPT-4o possesses perfect response inhibition ability." However, internal parameter analysis indicated that the model still experienced inhibitory conflict, while brain-model correlation analysis showed that this internal state lacks correspondence with human response inhibition networks. These three levels of evidence

collectively suggest that the model's "perfect performance" may not stem from response inhibition ability in the human sense, but from different computational mechanisms (35). Building on the work of Bauer et al. (2025), who proposed using log probability parameters to assess LLM uncertainty, in this study, we extended this method to cognitive psychology, exploring its potential applications in EF assessment (27). Future research could further investigate relationships between other internal parameters (e.g., attention weights, hidden layer activations) (43) and cognitive processing, thereby establishing a more comprehensive cognitive assessment toolkit.

Limitations and future considerations

Despite the meaningful findings, this study has several limitations that warrant consideration. First, we focused exclusively on GPT-4o as a model, similar to previous studies (44, 45). Different model architectures, scales, and training paradigms may yield distinct EF performance patterns (46). For instance, models incorporating more human feedback reinforcement learning might perform better at following "inhibition instructions," although whether this would constitute genuine "response inhibition ability" is unclear (47). Because the observed selective replication pattern may partially reflect specific characteristics of GPT series models, future research should assess multiple mainstream models to establish benchmark datasets for LLM EFs. Second, while this study covered three major EF dimensions based on the Miyake framework, the range of tasks within each component was limited. Comprehensive EF assessment necessitates a broader task repertoire, including Simon tasks, complex span tasks, the WCST, and others (48). Future research should expand the task range to achieve a more comprehensive characterization of LLM EF features. Third, in this study, while we undertook a preliminary exploration of log probability parameters as cognitive indicators, many unresolved questions remain. Regarding parameter selection and combination, Application Programming Interfaces (APIs) offer additional information such as top-*k* token probability distributions, and determining which parameter combinations best predict cognitive performance requires further investigation (49). Moreover, whether the psychological significance of log probability parameters remains consistent across different models also needs verification. Furthermore, this study primarily relied on correlational analysis; future work could attempt to manipulate log probability parameters and observe corresponding behavioral changes to explore potential causal relationships. Finally, the theoretical mechanisms underlying selective replication require deeper examination. Why can certain EF components be acquired through text learning while others cannot? Does this difference reflect evolutionary hierarchies of cognitive functions? Addressing these questions will enhance our understanding of the uniqueness of human cognition and inform future directions for artificial intelligence development.

Summary

In this study, we systematically evaluated GPT-4o's performance on EF tasks using standard cognitive paradigms employed in human studies, and found that the model selectively replicated human cognitive patterns. The model demonstrated performance comparable to humans in tasks based on symbolic representation and maintenance of static information, but exhibited divergent performance in tasks involving dynamic control and temporal sensitivity, as well as those requiring intrinsic variability. Log probability parameters, serving as quantitative indicators of model internal states, showed systematic associations with behavioral performance and human brain activation patterns in certain tasks, providing new analytical tools for assessing cognitive processing. This selective replication pattern not only clarifies the cognitive boundaries of current LLMs but also provides valuable empirical methods for exploring the computational foundations of EFs.

Materials and Methods

Participants

Two independent datasets were analyzed to investigate EF performance (SI Appendix, Table S2). Dataset 1 consisted of behavioral data collected for this study from 1,970 native Mandarin-speaking adults (915 males [46.45%], 1,055 females [53.55%]; mean age = 19.78 years, SD = 1.72) who completed five standardized EF tasks. Participants were recruited from Northwest Normal University and surrounding communities. The inclusion criteria comprised native Mandarin proficiency, normal or corrected-to-normal vision, and normal color vision. Individuals with neurological or psychiatric disorders or those taking medications affecting cognitive function were excluded. Dataset 2 consisted of fMRI data from He et al. (2021), involving 39 adults (13 males, 26 females; mean age = 19.21 ± 0.52 years) who performed three EF tasks during scanning (50). All human participants provided written informed consent, and the study was approved by the Ethics Committee of Northwest Normal University (Protocol #NUNU-20250105).

Testing human EF performance

Participants from Dataset 1 completed five EF tasks using E-Prime software (SI Appendix, Fig. S1; for details, see SI Appendix, Methods). The task set evaluated core EF components, and consisted of the following: (1) The *Stroop color-word* task, which assessed interference control by requiring participants to identify ink colors while ignoring word meanings across congruent, incongruent, and neutral conditions (18 practice trials, 108 test trials); (2) the *Go/No-Go* task, which measured response inhibition, with participants responding to target letters while withholding responses to non-targets in a counterbalanced design (4 blocks × 100 trials); (3) the *number-switching* task, which evaluated cognitive flexibility through alternating magnitude judgments for red digits and parity judgments for blue digits across single-task blocks (10 blocks × 8 trials) and mixed-task blocks (10 blocks × 17 trials); (4) the *running memory* task, which assessed working memory updating by requiring continuous updating of the last three digits under simple (1750 ms presentation) and difficult (750 ms) conditions across sequences of varying lengths (5–11 items); and (5) the *digit span* task, which measured working memory maintenance through forward and backward recall using adaptive procedures beginning with three-digit sequences.

Participants from Dataset 2 performed three tasks during fMRI scanning: (1) The *n*-back task, which assessed working memory across 0-back, 1-back, and 2-back conditions using blocked designs (4 blocks per condition, 29 s each with 15 stimuli); (2) the *stop-signal* task, which evaluated response inhibition with 96 Go trials and 32 stop trials, employing dynamically adjusted stop-signal delays; and (3) the *number-letter switching* task, which measured cognitive flexibility with 48 repeat and 48 switch trials, requiring alternation between parity and consonant/vowel judgments based on color cues.

Testing LLM EF performance and log parameters

GPT-4o (version 2025-03-26) was evaluated through OpenAI's official API with the temperature set to 0 to ensure consistency with prior LLM cognitive research (19, 20, 31, 44, 45), and thereby facilitate direct comparisons of results across studies. To verify the robustness of our findings, all analyses were repeated at temperature = 1 as a sensitivity check (20, 31). Default settings were used for all other parameters. The model completed EF tasks through prompt-driven multi-turn conversations, with each simulated participant maintaining consistent context within a single dialog session to ensure ecological validity and independence between participants. The task prompts and procedures closely matched those used with human participants (35), and were delivered to GPT-4o through natural language inputs (details in SI Appendix, Tables S3 and S4). To quantify the model's internal computational processes (SI Appendix, Table S5), four log probability parameters were extracted during response generation (see https://cookbook.openai.com/examples/using_logprobs):

(1) Average log probability quantified overall model confidence:

$$\text{avg\_logprob} = \frac{1}{n}\sum_{i=1}^{n} \log p\,(t_i)$$

(2) Perplexity measured model uncertainty:

$$\text{PPL} = \exp(-\text{avg\_logprob})$$

(3) Log probability variance indicated confidence variability:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (\log p\,(t_i) - \text{avg\_logprob})^2$$

(4) Minimum log probability identified weakest predictions:

$$\text{min\_logprob} = \min_{i=1,\ldots,n} \log p\,(t_i)$$

All computations were implemented using *NumPy* (51).

MRI data acquisition and analysis

fMRI data were acquired using a 3T Siemens scanner with echo-planar imaging sequences (TR = 2000 ms, TE = 30 ms, voxel size = 3.4 × 3.4 × 3 mm^3). High-resolution T1-weighted structural images were obtained using Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequences. Preprocessing followed standard procedures using Statistical Parametric Mapping 12 (SPM12) and Data Processing & Analysis for Brain Imaging (DPABI) toolboxes (52). These steps included slice-timing correction, realignment, coregistration, Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra (DARTEL) segmentation, normalization to Montreal Neurological Institute (MNI) space, and spatial smoothing with an 8-mm Full Width at Half Maximum (FWHM) Gaussian kernel. Task-based activation was modeled using general linear models with the following contrasts: “2-back > 0-back” for working memory load, “successful stop > successful Go” for inhibitory control, and “switch > repeat” for cognitive flexibility costs. Mean activation values were extracted from 116 anatomically-defined regions of interest using the Automated Anatomical Labeling (AAL) atlas (53) (see SI Appendix, Methods).

Sample size justification and considerations

Sample size determination was predicated on fundamental differences between human and LLM response distributions. While human samples inherently reflect genuine individual differences in EF, LLMs, when using fixed temperature parameters, generate statistical variation around a model-determined central tendency (20). Given the more homogeneous response patterns displayed by LLMs, excessive sampling introduces the risk of pseudo-precision and violates the value of information principle (54, 55). A total of 120 simulated participants were generated for Dataset 1, providing >90% power to detect medium effects (d = 0.5) and >99% power for large effects (d = 0.8) at α = 0.05, consistent with typical EF effect sizes (d = 0.4–0.8) (36). To ensure fair comparisons and mitigate statistical biases from unequal sample sizes, 120 participants from the original 1,970 human participants were randomly subsampled for matched analyses. For Dataset 2, 39 simulated participants were generated to match the fMRI sample size, enabling a direct comparison of internal computational parameters with neural activation patterns.

Data analysis

Bayesian ANOVA was used to compare EF performance (SI Appendix, Table S6) between AI and humans (56), with evidence strength quantified through BFs (57). Analyses were implemented in Python using Python Monte Carlo (PyMC3) with weakly informative priors (58). Posterior distributions for effect size differences were computed, and the resulting posterior means and medians were reported as point estimates along with 95% HDIs to characterize

uncertainty. BF served to quantify support for the hypothesis that AI effects exceed human effects:

$$BF_{10} = \frac{P(\Delta_{\text{effect}} > 0 \mid \text{data})}{P(\Delta_{\text{effect}} \leq 0 \mid \text{data})}$$

Following established conventions (59), BF_{10} >10 indicated strong evidence, 3< BF_{10} <10 indicated moderate evidence, 1< BF_{10} <3 indicated weak evidence, and BF_{10} <0.33 supported the null hypothesis.

To investigate the relationships between internal computational mechanisms and behavioral performance, Pearson correlation analyses were performed between task performance metrics (reaction times, accuracy) and log probability parameters. All correlation coefficients were corrected for multiple comparisons using the Benjamini-Hochberg False Discovery Rate (FDR) procedure at α = 0.05 (60). For brain-model correspondence analyses, we identified significantly activated ROIs through one-sample *t*-tests ($p_{Bonferroni}$ < 0.05), then performed Pearson correlations between ROI activation values and model log probability parameters. These analyses explored statistical associations between neural activity and model output uncertainty metrics, though such correlations should be interpreted as statistical relationships rather than evidence of shared computational mechanisms. All correlation coefficients were also corrected for multiple comparisons using FDR procedure at α = 0.05.

Data, materials, and software availability

The transcripts and the numerical coding of the transcripts have been deposited in the Open Science Framework (OSF) (https://osf.io/hfnt6).

Acknowledgments

We thank Peng Lu, Shuyan Gao, and Qiya Huang from the School of Psychology, Northwest Normal University, for their assistance with data collection; Yajie Gong and Jia Li from the School of Psychology, Northwest Normal University, for organizing the experimental materials; and all participants for their participation in this study. We also thank Nai Ding from the College of Biomedical Engineering and Instrument Sciences, Zhejiang University, for valuable insights on the research approach during the early stages of this study.

References

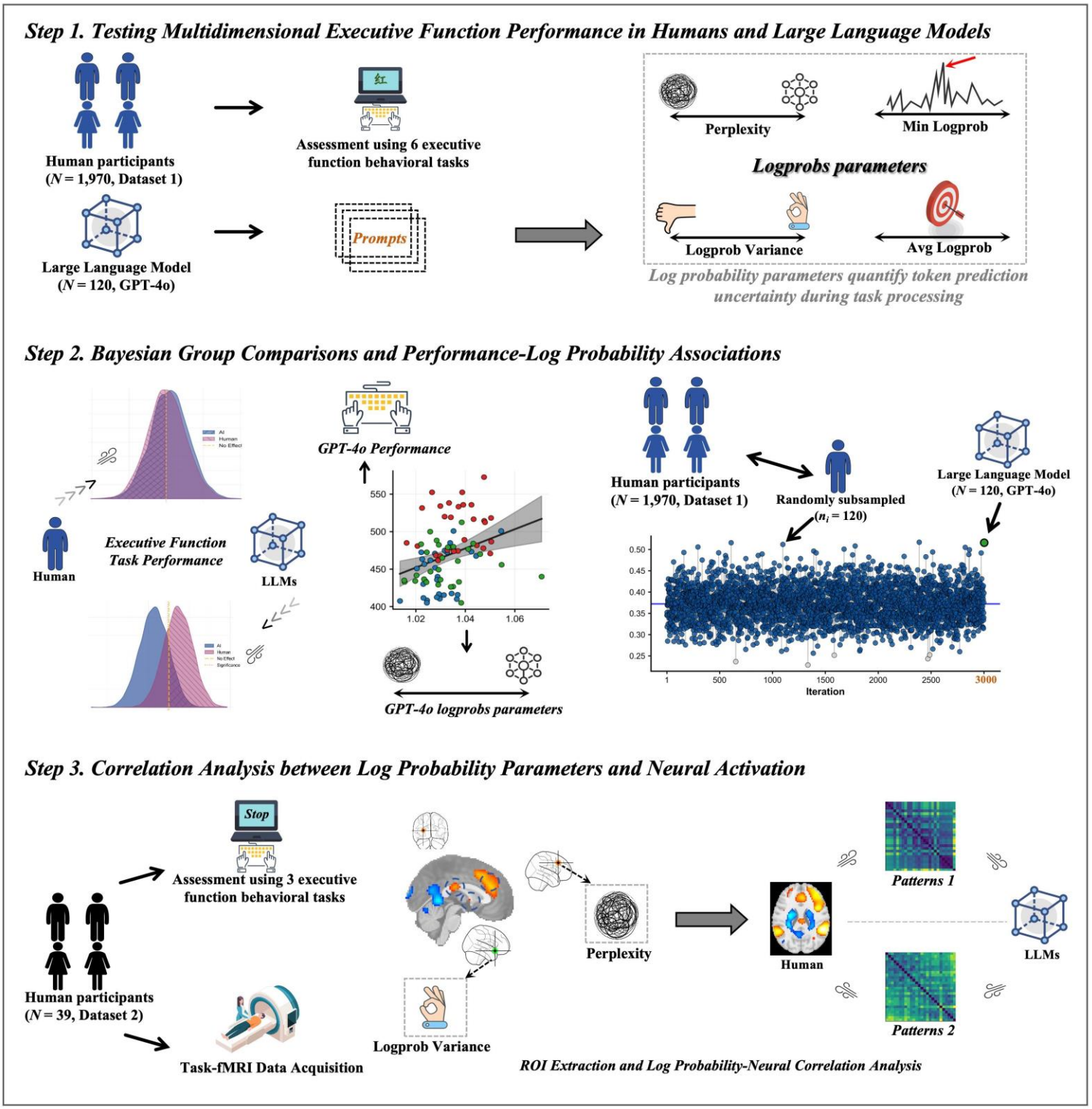
1. T. Hagendorff, S. Fabi, M. Kosinski, Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat Comput Sci* **3**, 833–838 (2023).
2. K. Valmeekam, A. Olmo, S. Sreedharan, S. Kambhampati, Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change) in (2022).
3. T. W. Webb, S. S. Mondal, C. Wang, B. Krabach, I. Momennejad, A Prefrontal Cortex-inspired Architecture for Planning in Large Language Models. *CoRR* (2023).
4. A. Chemero, LLMs differ from human cognition because they are not embodied. *Nat Hum Behav* **7**, 1828–1829 (2023).
5. M. C. Frank, Openly accessible LLMs can help us to understand human cognition. *Nat Hum Behav* **7**, 1825–1827 (2023).
6. L. Ke, S. Tong, P. Cheng, K. Peng, Exploring the frontiers of LLMs in psychological applications: a comprehensive review. *Artif Intell Rev* **58**, 305 (2025).
7. J. Meng, AI emerges as the frontier in behavioral science. *Proceedings of the National Academy of Sciences* **121**, e2401336121 (2024).
8. M. Hardy, I. Sucholutsky, B. Thompson, T. Griffiths, Large language models meet cognitive science: LLMs as tools, models, and participants. *Proceedings of the Annual Meeting of the Cognitive Science Society* **45** (2023).
9. K. Mahowald, *et al.*, Dissociating language and thought in large language models. *Trends in Cognitive Sciences* **28**, 517–540 (2024).

10. I. Yildirim, L. A. Paul, From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences* **28**, 404–415 (2024).
11. A. Diamond, Executive Functions. *Annu. Rev. Psychol.* **64**, 135–168 (2013).
12. N. P. Friedman, T. W. Robbins, The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacol.* **47**, 72–89 (2022).
13. M. Carlén, What constitutes the prefrontal cortex? *Science* **358**, 478–482 (2017).
14. D. Gong, X. Wan, D. Wang, Working Memory Capacity of ChatGPT: An Empirical Study. *AAAI* **38**, 10048–10056 (2024).
15. C. Zhang, Y. Jian, Z. Ouyang, S. Vosoughi, Working Memory Identifies Reasoning Limits in Language Models in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, Y.-N. Chen, Eds. (Association for Computational Linguistics, 2024), pp. 16896–16922.
16. X. Hu, R. L. Lewis, Do Language Models Understand the Cognitive Tasks Given to Them? Investigations with the N-Back Paradigm. [Preprint] (2024). Available at: http://arxiv.org/abs/2412.18120 [Accessed 29 December 2024].
17. K. de Langis, *et al.*, A Framework for Robust Cognitive Evaluation of LLMs. [Preprint] (2025). Available at: http://arxiv.org/abs/2504.02789 [Accessed 26 April 2025].
18. R. Loconte, G. Orrù, M. Tribastone, P. Pietrini, G. Sartori, Challenging ChatGPT ' *Intelligence*' with Human Tools: A Neuropsychological Investigation on Prefrontal Functioning of a Large Language Model. [Preprint] (2023). Available at: https://papers.ssrn.com/abstract=4377371 [Accessed 8 July 2025].
19. J. Coda-Forno, M. Binz, J. X. Wang, E. Schulz, CogBench: a large language model walks into a psychology lab. [Preprint] (2024). Available at: http://arxiv.org/abs/2402.18225 [Accessed 2 December 2024].
20. Z. Cui, N. Li, H. Zhou, A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nat Comput Sci* **5**, 627–634 (2025).
21. A. Miyake, N. P. Friedman, The nature and organization of individual differences in executive functions: Four general conclusions. *Current directions in psychological science* **21**, 8–14 (2012).
22. M. D'Esposito, B. R. Postle, The cognitive neuroscience of working memory. *Annual review of psychology* **66**, 115 (2015).
23. M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, J. D. Cohen, Conflict monitoring and cognitive control. *Psychological review* **108**, 624 (2001).
24. A. R. Aron, T. W. Robbins, R. A. Poldrack, Inhibition and the right inferior frontal cortex: one decade on. *Trends in Cognitive Sciences* **18**, 177–185 (2014).
25. M. Binz, *et al.*, A foundation model to predict and capture human cognition. *Nature* 1–8 (2025). https://doi.org/10.1038/s41586-025-09215-4.
26. H. Lin, Y. Zhang, Navigating the Risks of Using Large Language Models for Text Annotation in Social Science Research. *Social Science Computer Review* 08944393251366243 (2025). https://doi.org/10.1177/08944393251366243.
27. C. Bauer, L. T. Duc Dang, T. van den Beucken, J. Schuchhardt, R. Herwig, Systematic analysis of hepatotoxicity: combining literature mining and AI language models. *Front. Artif. Intell.* **8** (2025).
28. S. Chen, M. v. Maddali, C. Bluethgen, C. p. Langlotz, R. Raj, Leveraging Generative Pre-trained Transformer (GPT) Large Language Models (LLMs) For Interstitial Lung Diseases (ILD) Clinical Research. *Am J Respir Crit Care Med* **211**, A2086–A2086 (2025).
29. H. Gonen, S. Iyer, T. Blevins, N. A. Smith, L. Zettlemoyer, Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037* (2022).
30. M. Peeperkorn, T. Kouwenhoven, D. Brown, A. Jordanous, Is Temperature the Creativity Parameter of Large Language Models? [Preprint] (2024). Available at: http://arxiv.org/abs/2405.00492 [Accessed 30 October 2025].
31. M. Kosinski, Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences* **121**, e2405460121 (2024).
32. C. Li, Y. Qi, Toward accurate psychological simulations: Investigating LLMs' responses to personality and cultural variables. *Computers in Human Behavior* **170**, 108687 (2025).

33. J. Huang, K. Sun, W. Wang, M. Dredze, Language Models Do Not Have Human-Like Working Memory. [Preprint] (2025). Available at: http://arxiv.org/abs/2505.10571 [Accessed 31 October 2025].
34. D. Goto, H. Idei, Y. Shiozuka, T. Ogata, Performance of Large Language Models and Analysis of Responses in the Wisconsin Card Sorting Task in *2025 IEEE International Conference on Development and Learning (ICDL)*, (2025), pp. 1–7.
35. Q. Xu, *et al.*, Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nat Hum Behav* **9**, 1871–1886 (2025).
36. J. E. Karr, *et al.*, The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin* **144**, 1147–1185 (2018).
37. C. Gao, *et al.*, Increasing alignment of large language models with language processing in the human brain. *Nature computational science* 1–11 (2025).
38. M. M. Halassa, S. Kastner, Thalamic functions in distributed cognitive control. *Nat Neurosci* **20**, 1669–1679 (2017).
39. A. Doerig, *et al.*, High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence* 1–15 (2025).
40. A. Azaria, T. Mitchell, The Internal State of an LLM Knows When It's Lying. [Preprint] (2023). Available at: http://arxiv.org/abs/2304.13734 [Accessed 31 October 2025].
41. D. Haywood, F. D. Baughman, Multidimensionality in Executive Function Profiles in Schizophrenia: a Computational Approach Using the Wisconsin Card Sorting Task. *Comput Brain Behav* **4**, 381–394 (2021).
42. B. AlKhamissi, *et al.*, From Language to Cognition: How LLMs Outgrow the Human Language Network. [Preprint] (2025). Available at: http://arxiv.org/abs/2503.01830 [Accessed 31 October 2025].
43. A. Vaswani, *et al.*, Attention is all you need. *Advances in neural information processing systems* **30** (2017).
44. G. Suri, L. R. Slater, A. Ziaee, M. Nguyen, Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General* **153**, 1066–1075 (2024).
45. S. A. Lehr, K. S. Saichandran, E. Harmon-Jones, N. Vitali, M. R. Banaji, Kernels of selfhood: GPT-4o shows humanlike patterns of cognitive dissonance moderated by free choice. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2501823122 (2025).
46. X. Wang, *et al.*, CogLM: Tracking Cognitive Development of Large Language Models. [Preprint] (2025). Available at: http://arxiv.org/abs/2408.09150 [Accessed 31 October 2025].
47. S. Chaudhari, *et al.*, RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. *ACM Comput. Surv.* **58**, 53:1-53:37 (2025).
48. M. Yangüez, B. Bediou, J. Chanal, D. Bavelier, In search of better practice in executive functions assessment: Methodological issues and potential solutions. *Psychological Review* **131**, 402–430 (2024).
49. L. Shan, S. Luo, Z. Zhu, Y. Yuan, Y. Wu, Cognitive Memory in Large Language Models. [Preprint] (2025). Available at: http://arxiv.org/abs/2504.02441 [Accessed 30 October 2025].
50. L. He, *et al.*, Unity and diversity of neural representation in executive functions. *Journal of Experimental Psychology: General* **150**, 2193–2207 (2021).
51. C. R. Harris, *et al.*, Array programming with NumPy. *Nature* **585**, 357–362 (2020).
52. C.-G. Yan, X.-D. Wang, X.-N. Zuo, Y.-F. Zang, DPABI: data processing & analysis for (resting-state) brain imaging. *Neuroinformatics* **14**, 339–351 (2016).
53. E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, M. Joliot, Automated anatomical labelling atlas 3. *Neuroimage* **206**, 116189 (2020).
54. D. Lakens, Sample Size Justification. *Collabra: Psychology* **8**, 33267 (2022).
55. F. Anvari, D. Lakens, Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology* **96**, 104159 (2021).
56. E. J. Miller, *et al.*, AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological Science* **34**, 1390–1403 (2023).

57. J. K. Kruschke, T. M. Liddell, The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon Bull Rev* **25**, 178–206 (2018).
58. A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput* **27**, 1413–1432 (2017).
59. R. E. Kass, A. E. Raftery, Bayes factors. *Journal of the american statistical association* **90**, 773–795 (1995).
60. Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, I. Golani, Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research* **125**, 279–284 (2001).

Figures



727 **Figure 1.** Study design and analytical framework. This study comprised three main steps. Step 1:
728 Data collection. We collected behavioral data from 1,970 human participants who completed five
729 standardized executive function (EF) tasks (Dataset 1). Concurrently, we tested GPT-4o (*N* =

120) on the same tasks through prompt-driven multi-turn dialogs, extracting log probability parameters (perplexity, log probability variance, average log probability, and minimum log probability) as indicators of the model's internal computational states. Step 2: Behavioral analysis. We compared EF task performance between GPT-4o and humans using Bayesian hierarchical ANOVA to examine canonical cognitive effects (e.g., Stroop interference, task-switching costs). To ensure statistical robustness, we randomly subsampled 120 human participants to match the GPT-4o sample size, repeating this procedure 3,000 times. Effect size differences were quantified using Bayes factors. We additionally examined correlations between model log probability parameters and behavioral performance metrics (FDR-corrected). Step 3: Neural correspondence analysis. We analyzed fMRI data from 39 participants (Dataset 2) who performed three EF tasks (*n*-back, *stop-signal*, and *number-letter switching*). Task-related activation was modeled using general linear models with predefined contrasts. Mean activation values were extracted from 116 brain regions defined by the Automated Anatomical Labeling (AAL) atlas, and correlations with GPT-4o log probability parameters were assessed for regions showing significant task-related activation.

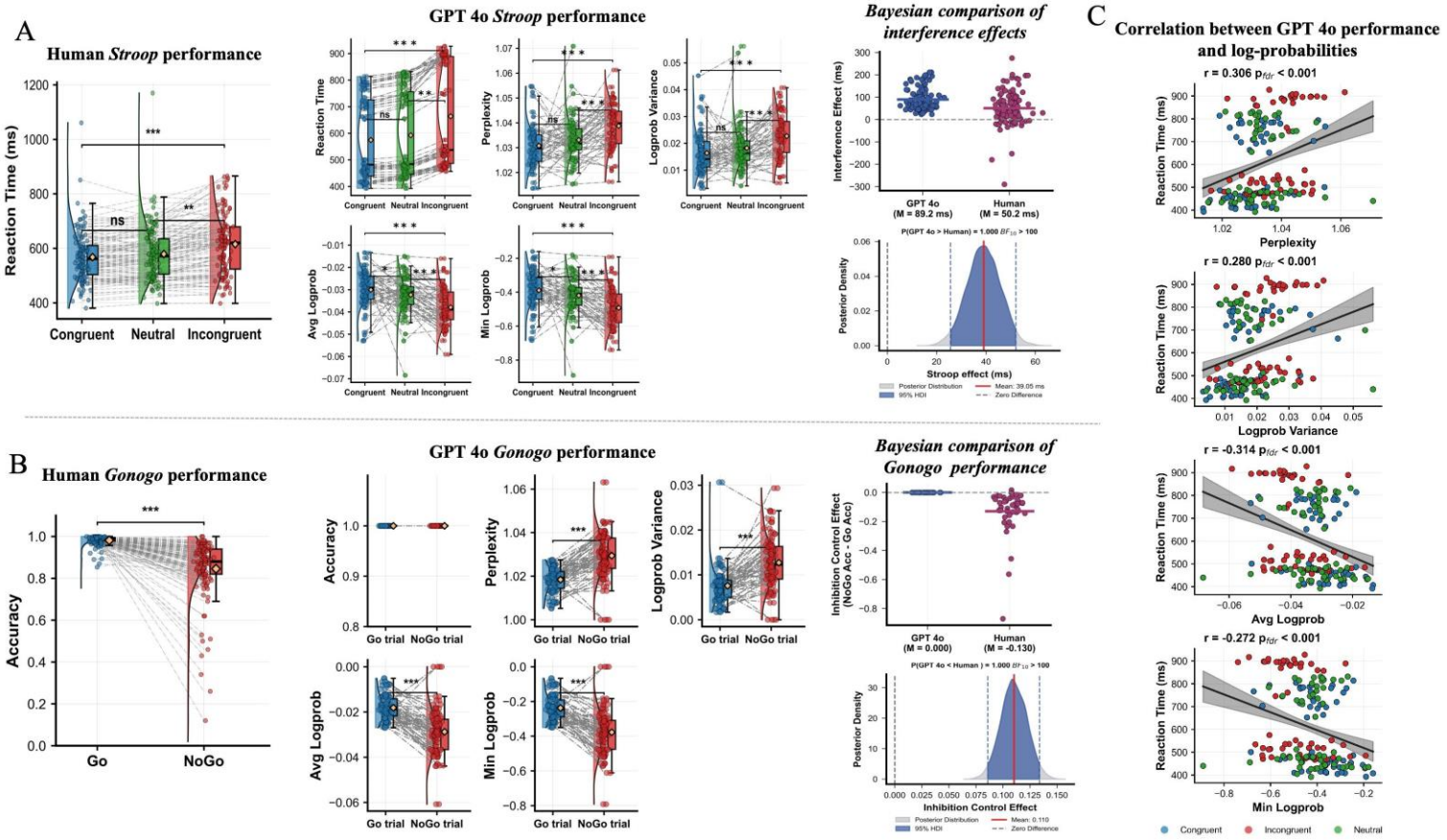


Figure 2. Behavioral performance comparison between GPT-4o and human participants in *inhibitory control* tasks. (A) Reaction times across three *Stroop* task conditions for human participants (*N* = 120) and the corresponding reaction times with log probability parameters for GPT-4o. The rightmost panel shows Bayesian analysis comparing interference effects (incongruent minus congruent condition) between humans and GPT-4o. (B) Accuracy in Go *versus* No-Go trials for human participants (*N* = 120) and the corresponding accuracy with log probability parameters for GPT-4o in the *Go/No-Go* task. The rightmost panel presents Bayesian analysis comparing inhibitory control effects (No-Go accuracy minus Go accuracy) between humans and GPT-4o. (C) Correlation analysis between reaction times and log probability parameters across the three *Stroop* task conditions for GPT-4o.

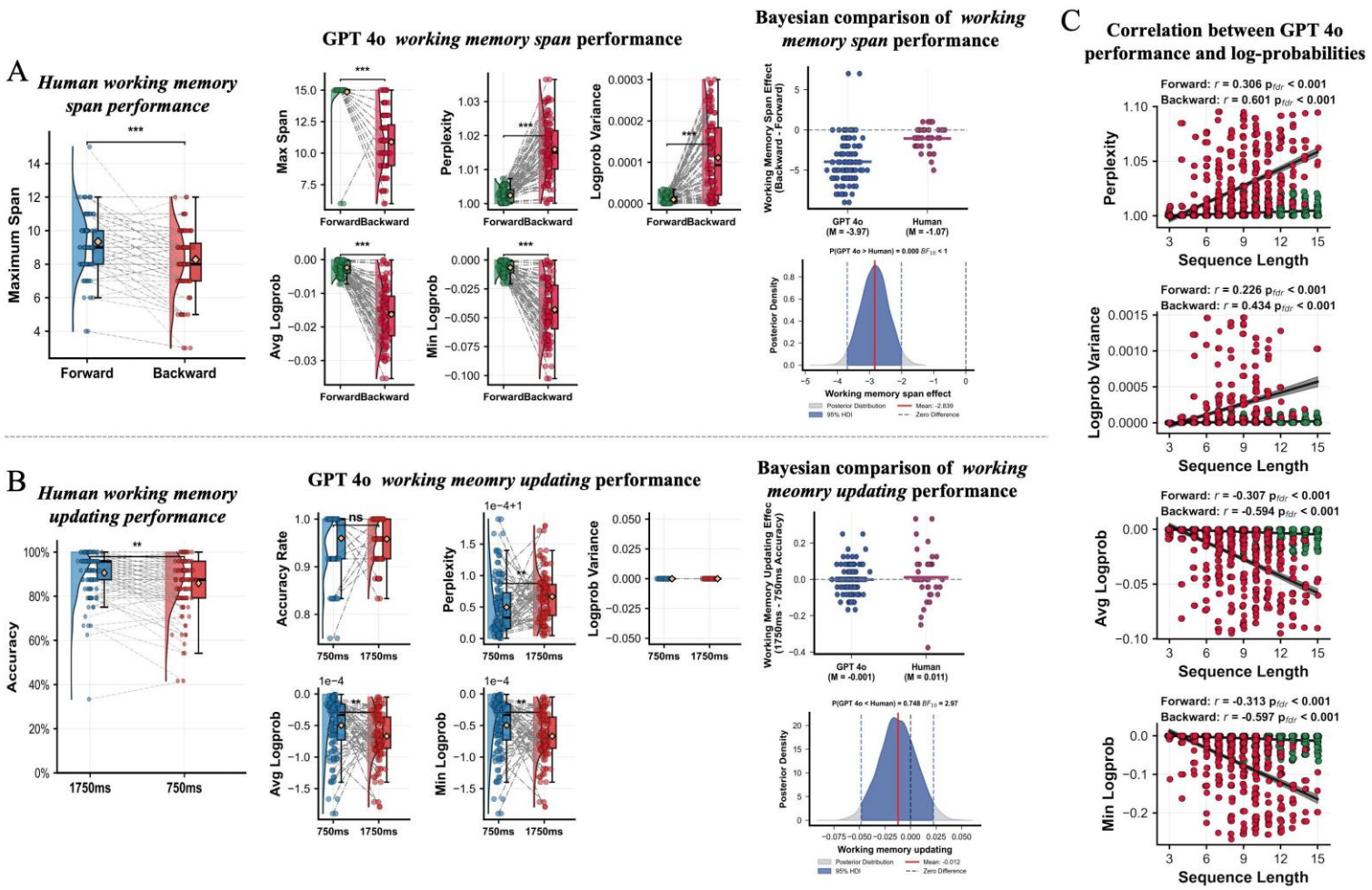


Figure 3. Performance comparison between GPT-4o and human participants in working memory tasks. (A) Working memory span in forward *versus* backward *digit span* tasks for human participants (*N* = 120) and the corresponding working memory span with log probability parameters for GPT-4o. (B) Accuracy at different stimulus presentation times (1750 and 750 ms) in the running digit memory task for human participants (*N* = 120) and the corresponding accuracy with log probability parameters for GPT-4o. The rightmost panel shows Bayesian analysis comparing refresh time effects (accuracy at 750 ms minus accuracy at 1750 ms) between humans and GPT-4o. (C) Bayesian analysis comparing working memory load effects (forward span minus backward span) between humans and GPT-4o. (D) Correlation analysis between sequence length and log probability parameters in forward and backward digit span tasks for GPT-4o.

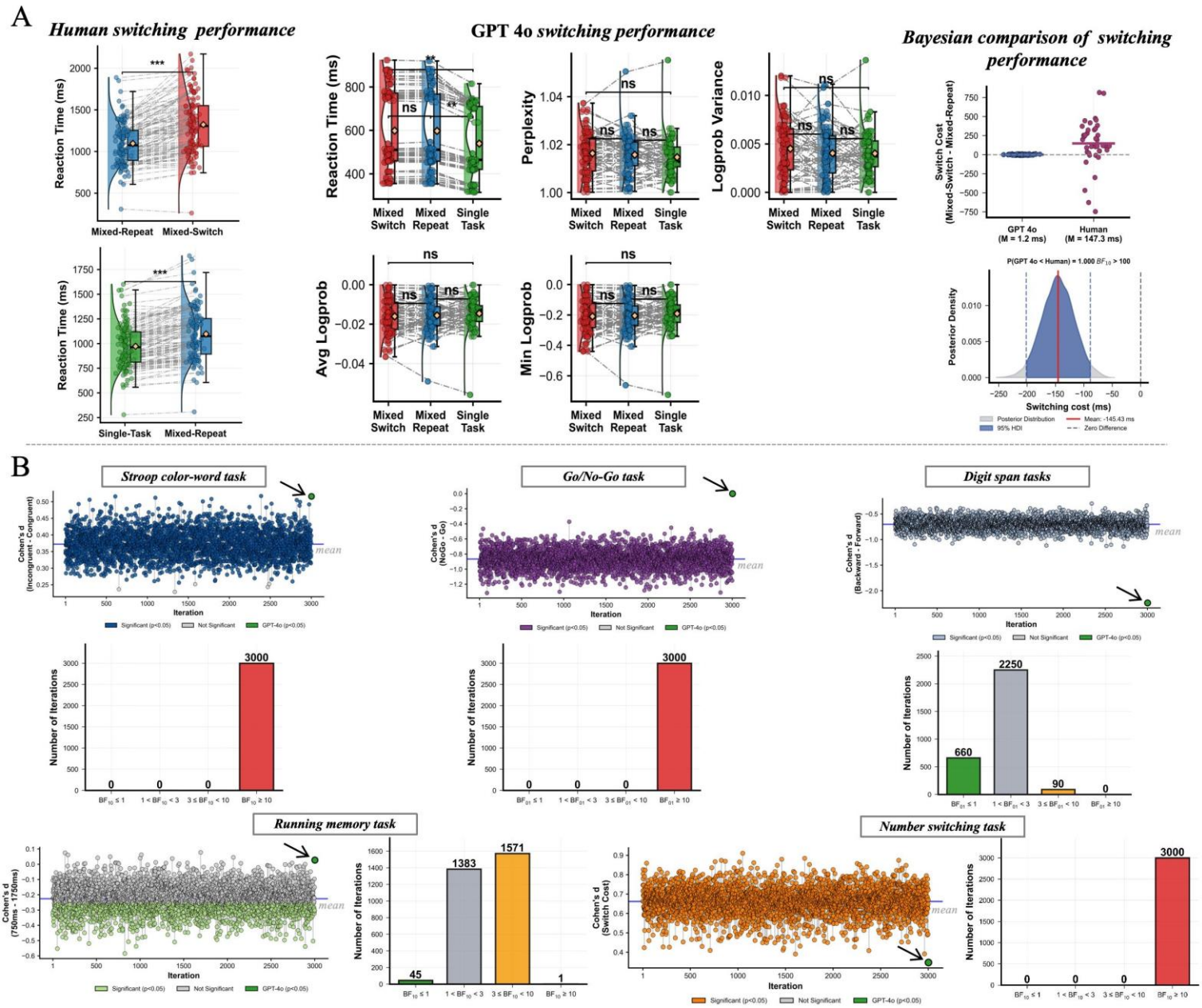


Figure 4. Comparison between GPT-4o and human participants in cognitive flexibility tasks and subsampling analysis. (A) Switching costs and mixing costs for human participants (N = 120) in the *number-switching* task, with corresponding reaction times and log probability parameters across three task conditions for GPT-4o. The rightmost panel presents a Bayesian analysis comparing switching costs between humans and GPT-4o. (B) Distribution of results from 3,000 repeated random samples (N = 120 each) drawn from the full human dataset (N = 1,970), demonstrating the statistical robustness of the primary findings.

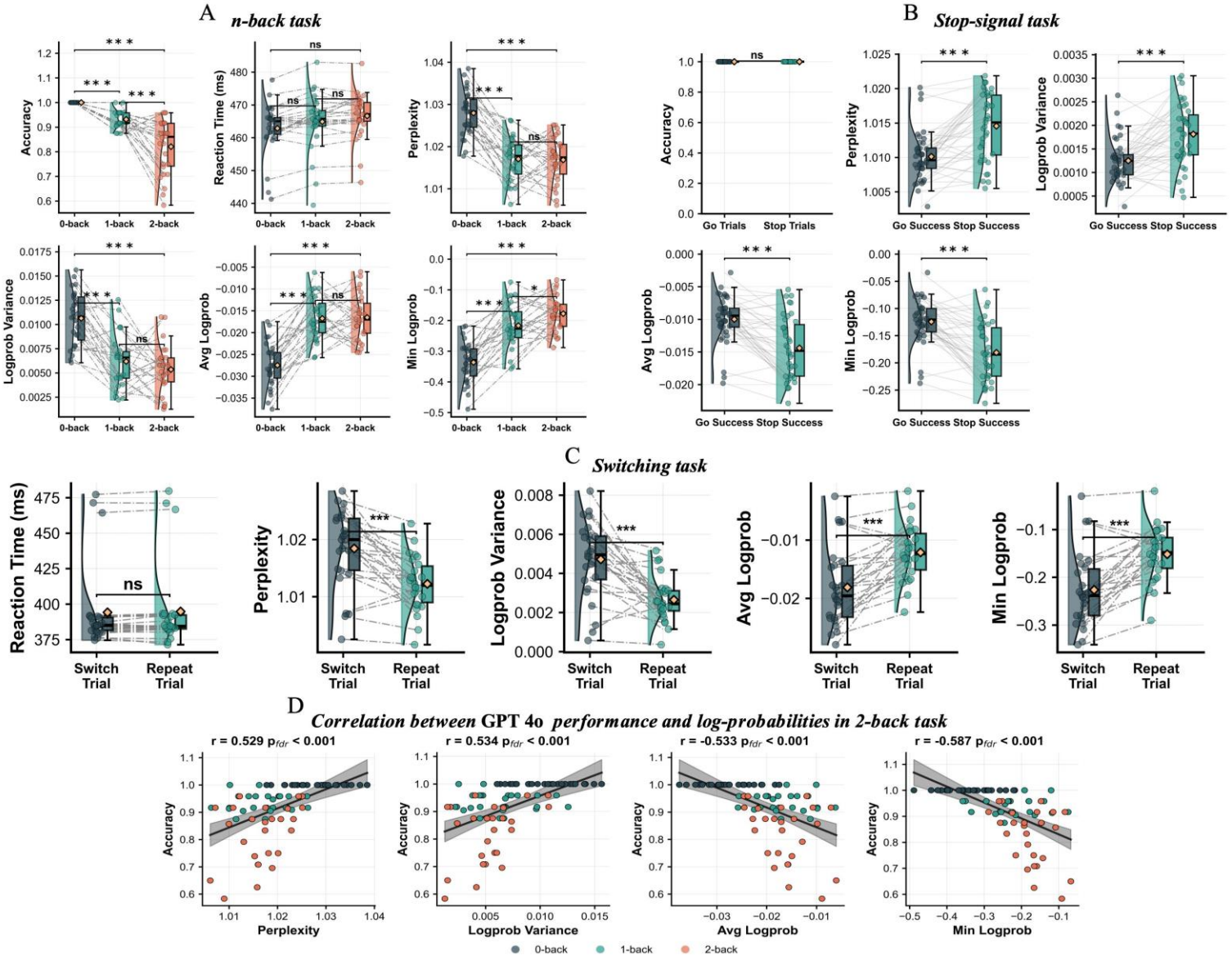


Figure 5. Behavioral performance in three executive function tasks from independent samples. (A) Performance comparison between human participants and GPT-4o in the *n*-back task. (B) Performance comparison between human participants and GPT-4o in the *stop-signal* task. (C) Performance comparison between human participants and GPT-4o in the *number-letter category-switching* task. (D) Correlation analysis between behavioral performance metrics and log probability parameters in the 2-back task for GPT-4o.

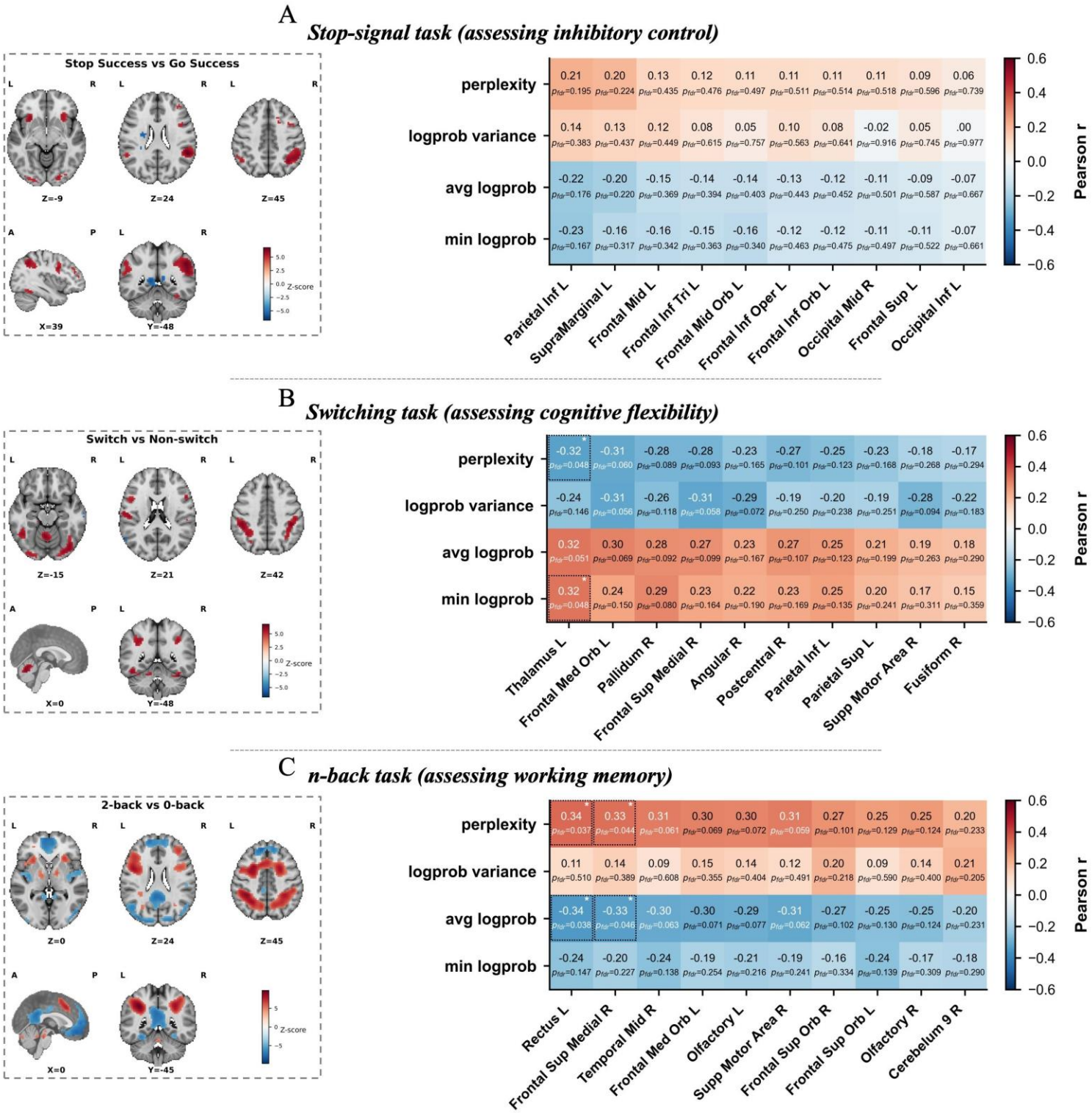


Figure 6. Relationship between GPT-4o internal log probability parameters and human brain neural activity. (A) *Stop-signal* task (measuring response inhibition). Left panel displays group-level brain activation patterns for the "successful stop > successful Go" contrast (Bonferroni corrected, $p < 0.05$); right panel presents correlation analyses between mean activation values

from significantly activated brain regions (116 ROIs defined by the Automated Anatomical Labeling [AAL] atlas) and GPT-4o log probability parameters. (B) *Number-letter category-switching* task (measuring cognitive flexibility). The left panel shows brain activation patterns for the "successful switch > successful repeat" contrast; the right panel displays correlation analyses between significantly activated regions and model parameters. (C) *n*-Back task (measuring working memory). The left panel illustrates brain activation patterns for the "2-back > 0-back" contrast; the right panel shows correlation analyses between significantly activated regions and model parameters.