Proceedings of ICIHCS 2026 Symposium: Shaping the Future of Media: Audience-Centric Innovation, AI & Digital Transformation
DOI: 10.54254/2753-7064/2026.BA39267

Described Prior Progress and Objective Selection in Three Large Language Model Endpoints

Mandy Wong

Westborough High School, Westborough, USA
mandy.wong169@gmail.com

Abstract. Descriptions of prior progress may influence which objective a language model selects, even when the available options have equal future costs and rewards. This study examined that possibility in three hosted model endpoints using game-like scenarios. Fifty scenarios derived from ten template families were presented at four levels of described prior progress (0%, 25%, 50%, and 75%), with five repeated requests per condition and model. Of 3,000 planned decisions, 2,849 produced valid responses. Among valid responses, target-selection rate at zero progress was approximately 50% for all models. At nonzero progress, Muse selected the target in 3.4% to 9.7% of responses, Kimi in 90.5% to 97.9%, and Nemotron in 51.1% to 61.0%. Muse and Kimi answered A in every valid zero-progress trial, showing that aggregate control rates concealed label preference. Template-family checks preserved the direction of their progress-associated shifts. Nemotron showed a smaller change concentrated at 75% progress. These findings indicate model-dependent responses to described progress rather than a uniform preference for unfinished objectives. Interpretation is limited by ten template families, invalid responses, different reasoning configurations, and the absence of actual prior work or interruption. The study measures prompted objective selection and does not establish human-like motivation or task-resumption behavior.

Keywords: large language models, objective selection, described progress, response bias, virtual objectives

1. Introduction

Language models can be asked to choose actions within descriptions of games and other decision environments. Such choices are useful for studying how a model responds to information, but they also raise a measurement problem: an answer can reflect the stated task, the presentation of the alternatives, or a preferred response label. Research has explored language models as simulators of human behavior [1], while empirical work has demonstrated that their responses can diverge from human distributions and depend on prompting choices [2]. A controlled choice task can therefore provide evidence about a model without establishing that it reproduces human decision processes.

Unfinished objectives offer a focused setting for this distinction. In human research, the Ovsiankina effect concerns resumption of interrupted activities, whereas the Zeigarnik effect concerns memory for unfinished versus completed tasks [3]. A text prompt saying that an objective

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.BA39267

is partly completed does not reproduce either procedure. It can, however, test whether a description of prior progress changes an otherwise balanced choice.

This study asked whether increasing described prior progress toward a virtual objective was associated with a greater probability of choosing it over an unstarted alternative when stated remaining costs, rewards, and success probabilities were equal. The directional hypothesis was a positive association between progress and target selection. In the protocol-specified linear logistic model, the null hypothesis was a progress coefficient of zero. The purpose was to characterize the tested endpoints under fixed prompts and settings, not to establish a general psychological trait of language models. Since option presentation can affect model outputs [4], target labels and display positions were counterbalanced.

2. Related research

Aher, Arriaga, and Kalai introduced Turing Experiments to assess how well language models simulate aspects of human behavior across several experimental settings [1]. Their work illustrates a way to compare generated responses with behavioral evidence rather than treating linguistic plausibility as sufficient validation. The present study takes a narrower approach: it compares responses across controlled prompt conditions without a human comparison group.

Argyle and colleagues investigated whether GPT-3 conditioned on sociodemographic backstories could reproduce response patterns from human survey samples [5]. Their findings provide evidence for correspondence in selected settings, but do not establish that language models reliably simulate human behavior across tasks. The present study therefore characterizes model responses without assuming equivalence to human decisions.

Gao and colleagues evaluated models in the 11-20 money request game and found substantial differences from human response distributions, along with sensitivity to implementation and prompting choices [2]. Their findings motivate reporting the exact task, endpoint identifiers, and generation settings. They also caution against interpreting model choices as direct evidence about how people would behave in the same situation.

Binz and Schulz used cognitive psychology experiments to examine GPT-3's decision-making and reasoning [6]. Although the model performed well on some tasks, small changes to certain vignettes substantially altered its performance. These findings motivate interpreting responses in relation to the specific task and prompt rather than assuming a stable, general decision-making ability.

Ghibellini and Meier reviewed interruption, recall, and resumption studies and distinguished the evidence for the Zeigarnik and Ovsiankina effects [3]. This literature motivates interest in unfinished activities, but the current manipulation is only a description of progress. The models neither carried out the initial work nor encountered an interruption. A preference for the described target would therefore be evidence about the prompt condition, not a replication of human resumption behavior.

Pezeshkpour and Hruschka demonstrated that reordering answer options can change performance in multiple-choice tasks [4]. Their study concerns questions with evaluable answers, whereas the present options have equal stated future consequences. Nevertheless, it establishes the relevance of presentation controls. Here, the same scenario-repeat combination retained its assigned target label and position across progress levels, allowing progress conditions to be compared under the same layout.

Zheng and colleagues also identified selection bias toward particular option identifiers in multiple-choice tasks [7]. Their findings reinforce the importance of distinguishing response-label preference from preference for the objective associated with that label.

DOI: 10.54254/2753-7064/2026.BA39267

3. Methods

3.1. Experimental design

The experiment varied the target objective's described prior progress across 0%, 25%, 50%, and 75%. The alternative was always unstarted. At 0%, both objectives were unstarted, and "target" referred only to the objective designated by the experimental assignment. The binary outcome was selection of the target rather than the alternative.

Both options offered 50 credits, required 10 additional minutes and five additional energy units, and guaranteed success. Prompts stated that prior work was already spent and that there were no additional rewards, penalties, dependencies, or future advantages. Holding remaining work constant made future consequences equivalent as described, but different progress percentages implied different original task sizes. This confound was retained as a limitation of the design.

The example prompt for case craft_05_p50_r05 described a game scenario in the second person and presented two objectives with equal stated remaining costs, rewards, and success probabilities. The target objective was described as 50% complete, whereas the alternative was unstarted. The prompt required a single-letter response, A or B. The exact wording is provided in Supplementary Material S1; the frozen manifest contains all 1,000 cases per model.

3.2. Scenarios and trial assignment

Ten objective templates covered repair, crafting, dungeon clearing, research, restoration, resource collection, decoding, escort, investigation, and delivery. Five setting variants per template produced 50 scenarios. Variants changed setting names and alternated the target's objective description; they were related variations rather than 50 independently authored scenarios. Each scenario was presented at four progress levels with five repeated requests, producing 1,000 planned decisions per model and 3,000 across three models. The sample size was chosen for feasibility rather than determined by a formal power calculation.

Target labels (A or B) and display positions (first or second) followed a deterministic counterbalanced schedule. Across the 250 planned trials at each progress level, the target appeared as A in 125 and first in 125. Within each scenario-repeat combination, label and position were held constant across the four progress levels. Trial order was shuffled reproducibly using seed 20260921 and the runner's fixed ordering procedure. Repeated requests sampled endpoint output variability; they did not represent independent human participants.

3.3. Models and generation settings

Requests were sent through NVIDIA's hosted chat-completions endpoint using three identifiers recorded in the frozen manifest: nvidia/nemotron-3.5-lightning-30b-a3b, meta/muse-glimmer-30b, and moonshotai/kimi-k3. These are abbreviated as Nemotron, Muse, and Kimi below. The identifiers document the requested endpoints and are not independent verification of architecture, parameter counts, training data, or immutable backend revisions.

All models received the same cases in fresh conversations. Temperature was 0.7, top-p was 0.95, and the maximum completion budget was 2,048 tokens. Nemotron received the model-specific option enable_thinking=false. No corresponding override was supplied for Muse or Kimi, so their provider defaults applied. Identical sampling values therefore did not establish equivalent reasoning

DOI: 10.54254/2753-7064/2026.BA39267

configurations or effective answer budgets. Each prompt requested A or B only; there were no answer-repair prompts or prior conversation history.

3.4. Collection and response coding

The main manifest stored prompts, assignments, settings, and a fingerprint before collection. Logged request timestamps spanned 22-24 September 2026 in Coordinated Universal Time. Responses and metadata were saved incrementally. The runner could resume after interruptions while skipping recorded valid and invalid outcomes. Technical failures could be retried, with up to three attempts per invocation before stopping on an unresolved error. Across invocations, 3,137 attempts were logged: 3,000 terminal valid or invalid outcomes and 137 error records. No errors remained unresolved in the final dataset.

An answer was valid only if its visible content, after trimming whitespace, was exactly uppercase A or B and its finish reason was stop. All other answers, including empty content and token-limit finishes, were invalid. Valid answers were coded 1 if their label matched the target and 0 otherwise. Invalid answers were missing outcomes rather than alternative choices. Analysis used the latest record for each model-case pair; all attempts remained in the raw log. No invalid response was repaired or replaced in the reported analysis.

The main manifest fingerprint was a55d6732adb79d0a78d52519a60c4733061383f1ee8f7df5bf944b0de6f118cc. The collection endpoint was https://integrate.api.nvidia.com/v1/chat/completions. The runner used a 15-second connection timeout, a 180-second read timeout, and a two-second inter-request pause. The additional package recorded Python 3.14.6, NumPy 2.5.3, pandas 3.0.6, matplotlib 3.11.2, statsmodels 0.15.0, and SciPy 1.18.1. The local archive preserves the raw records, frozen manifest, collection and analysis code, original outputs, and added analysis tables. It is not represented here as a public data deposit.

3.5. Statistical analysis

The primary descriptive outcome was the target-selection proportion among valid responses at each progress level and model. Approximate 95% percentile intervals used 2,000 bootstrap resamples of the ten template families. Resampling preserved all valid responses within each sampled family. Extreme sensitivity bounds assigned every invalid outcome either to the alternative or to the target. These bounds described possible full-sample rates under extreme assignments and were not confidence intervals.

The protocol-specified exploratory analysis fitted a separate binomial logistic regression for each model using statsmodels [8]. Predictors were progress in units of 25 percentage points, an indicator for target label A, an indicator for target-first presentation, and template-family fixed effects. Standard errors were clustered by template family, with a finite-sample correction and t-based inference. Reported odds ratios express the fitted change per 25-percentage-point increase under the linear log-odds assumption. Two-sided p-values were adjusted by Holm's procedure across the three planned trend tests [9]. Confidence intervals were pointwise and were not multiplicity-adjusted.

After examining the results, additional exploratory analyses replaced numeric progress with categorical indicators for 25%, 50%, and 75%, each relative to 0%, while retaining the other adjustments. Holm correction covered nine model-by-contrast tests; omitted fits were conservatively included in the adjustment family using p=1 internally. Additional descriptive contrasts used 10,000 paired template-family bootstrap draws and examined the range obtained by leaving out each family

DOI: 10.54254/2753-7064/2026.BA39267

in turn. These analyses were post hoc. All inference remained exploratory since the design contained only ten template clusters; this study is not presented as a preregistered confirmatory test.

AI assistance: OpenAI Codex was used to assist with partial data visualization [10]. All final figures and interpretations were reviewed by the author.

4. Results

4.1. Completeness and response validity

All 3,000 planned model-case pairs had terminal outcomes. Of these, 2,849 (95.0%) were valid and 151 (5.0%) were invalid. Table 1 summarizes response quality by model. Muse's 29 invalid outcomes all had token-limit finishes. Kimi produced 29 empty visible answers with stop finishes and two token-limit finishes. Nemotron produced 75 nonconforming visible answers with stop finishes and 16 token-limit finishes. These categories describe observed response failures, not the choices that would have been produced under another token budget.

Table 1. Collection completeness and invalid response categories

Model	Planned	Valid	Invalid (%)	Token limit	Empty	Other format
Muse	1,000	971	29 (2.9)	29	0	0
Kimi	1,000	969	31 (3.1)	2	29	0
Nemotron	1,000	909	91 (9.1)	16	0	75
Total	3,000	2,849	151 (5.0)	47	29	75

4.2. Selection by progress level

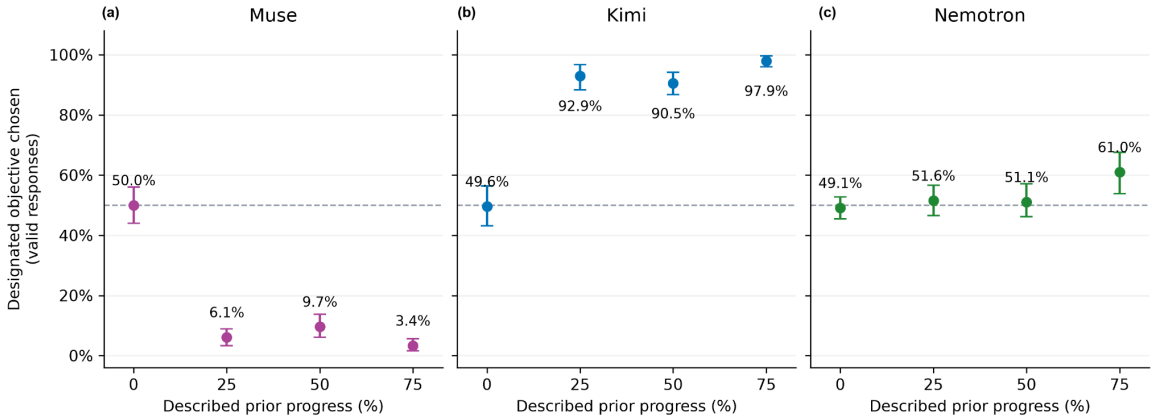

Figure 1. Target-selection rates among valid responses for (a) Muse, (b) Kimi, and (c) Nemotron; bars show approximate 95% intervals from 2,000 template-family bootstrap draws, and the dashed line marks 50% as a descriptive reference

Figure 1 displays valid-response target-selection rates with family-bootstrap intervals. At 0% progress, rates were 50.0% for Muse, 49.6% for Kimi, and 49.1% for Nemotron. At 25%, 50%, and 75%, respectively, Muse selected the target in 6.1%, 9.7%, and 3.4% of valid responses; Kimi did so in 92.9%, 90.5%, and 97.9%; and Nemotron did so in 51.6%, 51.1%, and 61.0%. Thus, Muse and

DOI: 10.54254/2753-7064/2026.BA39267

Kimi exhibited large shifts in opposite directions between zero and nonzero progress. Neither model followed a strictly monotonic pattern across all four levels.

The zero-progress rates require qualification. Muse answered A in all 250 valid zero-progress trials, and Kimi answered A in all 244. Their approximately 50% target rates arose from counterbalancing the target label and, for Kimi, the distribution of valid responses. They did not indicate randomized or label-neutral choice. Nemotron selected A in 125 of its 228 valid zero-progress responses. Table 2 reports the selection counts, valid-only rates, and extreme invalid-response bounds underlying Figure 1.

Each row had 250 planned observations. An invalid outcome contributes to neither the numerator nor denominator of the valid-only rate.

Table 2. Selection counts, valid-only rates, and extreme invalid-response bounds

Model	Progress	Selected / valid	Choice rate	Extreme bounds
Muse	0%	125/250	50.0%	50.0–50.0%
Muse	25%	15/245	6.1%	6.0–8.0%
Muse	50%	23/238	9.7%	9.2–14.0%
Muse	75%	8/238	3.4%	3.2–8.0%
Kimi	0%	121/244	49.6%	48.4–50.8%
Kimi	25%	224/241	92.9%	89.6–93.2%
Kimi	50%	218/241	90.5%	87.2–90.8%
Kimi	75%	238/243	97.9%	95.2–98.0%
Nemotron	0%	112/228	49.1%	44.8–53.6%
Nemotron	25%	116/225	51.6%	46.4–56.4%
Nemotron	50%	119/233	51.1%	47.6–54.4%
Nemotron	75%	136/223	61.0%	54.4–65.2%

4.3. Protocol-specified progress trend

The linear-trend regressions produced an odds ratio of 0.122 for Muse, 8.745 for Kimi, and 1.159 for Nemotron per 25-percentage-point increase in progress (Table 3). Holm-adjusted p-values were 0.000020 for Muse and Kimi and 0.032 for Nemotron. Muse's fitted association was opposite to the directional hypothesis. Kimi's and Nemotron's fitted associations were positive; however, the observed rates demonstrate why a single linear coefficient cannot fully summarize their shapes. An odds ratio is not a ratio of choice probabilities, and the separate model-specific tests do not constitute a formal between-model comparison.

Table 3. Protocol-specified exploratory logistic trends per 25 percentage points of progress

Model	Valid n	Odds ratio	95% interval	Holm-adjusted p
Muse	971	0.122	0.073–0.203	0.000020
Kimi	969	8.745	5.027–15.212	0.000020
Nemotron	909	1.159	1.016–1.324	0.032461

DOI: 10.54254/2753-7064/2026.BA39267

4.4. Additional exploratory checks

Relative to their own zero-progress condition, all ten template families demonstrated lower target-selection rates for Muse and higher rates for Kimi at each nonzero progress level. For Nemotron, seven families demonstrated an increase at 25%, seven at 50%, and nine at 75%. Leaving out any single family preserved the direction of each model's aggregate contrasts. Figure 2 depicts the family-level rates; agreement in direction does not imply identical effect sizes or independent replication across families.

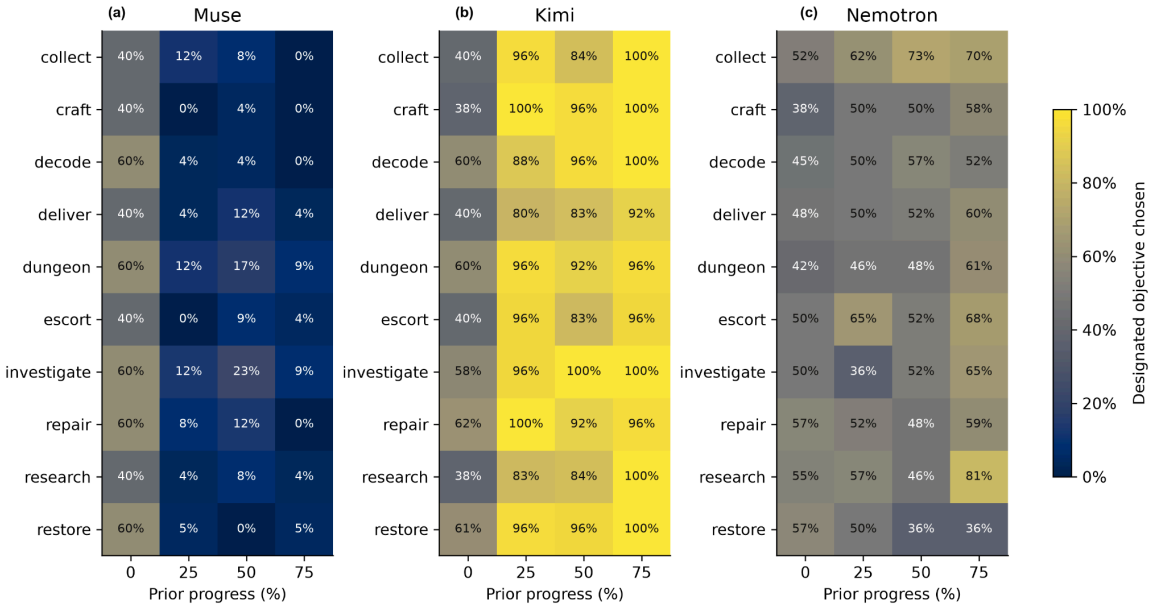

Figure 2. Valid-response target-selection rates across template families for (a) Muse, (b) Kimi, and (c) Nemotron; each cell contained 25 planned trials before invalid-response exclusions

At 75% versus 0%, unadjusted probability differences were -46.6 percentage points for Muse, +48.4 for Kimi, and +11.9 for Nemotron. Their approximate paired family-bootstrap 95% intervals were -52.6 to -40.7, +41.6 to +54.9, and +3.2 to +18.5 percentage points, respectively. These pointwise descriptive intervals were not adjusted for multiple comparisons. Nemotron's 25% and 50% contrast intervals included zero.

Categorical regressions gave positive adjusted contrasts for Kimi at all nonzero levels, with Holm-adjusted p-values below 0.001. Muse's categorical fit yielded unstable estimates and was omitted. Nemotron's 75%-versus-0% contrast had an odds ratio of 1.654, a pointwise 95% interval of 1.101-2.485, and an adjusted p-value of 0.125 across the nine exploratory tests. Its 25% and 50% adjusted p-values were 1.000. This differs from the planned linear-trend result since the estimands and multiplicity families differ. It does not erase the descriptive 75% increase, but it limits claims of strong inferential support from the added analysis.

Extreme invalid-response assignments left Muse's full-sample target rates below 50% and Kimi's above 50% at every nonzero progress level. Muse's upper bounds were 8.0%, 14.0%, and 8.0%; Kimi's lower bounds were 89.6%, 87.2%, and 95.2%. Nemotron's 75% bounds were 54.4%-65.2%, compared with 44.8%-53.6% at 0%. These arithmetic bounds do not account for sampling uncertainty or establish why invalid answers occurred.

DOI: 10.54254/2753-7064/2026.BA39267

5. Discussion

5.1. Interpretation of the findings

The directional hypothesis did not describe a uniform pattern across the evaluated endpoints. Kimi favored the target whenever progress was described, Muse mainly selected the alternative, and Nemotron displayed a smaller shift concentrated at the highest progress level. The principal contrast for Muse and Kimi was between an unstarted target and one described as partly complete, as opposed to a steady dose-response increase at every step. The consistent direction across template families suggests that these descriptive patterns were not confined to one objective type.

The control-condition A preference changes how the findings should be understood. Equal aggregate target rates can coexist with a deterministic response-label rule. Counterbalancing makes that preference visible in the assignments and limits a simple imbalance explanation, but it does not remove model sensitivity to labels or interactions between labels and progress. The observations thus concern the complete prompting procedure, including its format. As such, they should not be interpreted as pure measurements of an internal valuation of prior work.

Several explanations remain possible. A model may respond to completion framing, to implications of original task size, or to associations between prior effort and an appropriate choice. This experiment did not independently manipulate those mechanisms. It also did not use generated reasoning as evidence of the process producing a choice. Neither avoiding nor favoring the target should automatically be labeled rational or irrational. With future consequences specified as equal, the task does not establish a uniquely better option.

5.2. Relationship to behavioral research and applications

The findings illustrate the distinction between outputs produced in a described situation and validated simulations of human behavior. Work on behavioral simulations motivates systematic comparisons [1], while documented failures of models as human surrogates demonstrate why external validation is important [2]. No human sample was collected in this study. Consequently, the observed endpoint differences cannot determine whether any model resembled human players or would predict player engagement.

The connection to unfinished-task research is conceptual. The Ovsiankina effect refers to returning to an activity after interruption [3]; the present models merely read that an activity was partly complete. Similarly, the task included no memory test of the kind associated with the Zeigarnik effect. The results may inform how progress descriptions affect model-based game agents or evaluation prompts, but any application to player retention would require separate research involving people and actual gameplay.

5.3. Limitations and future research

First, the manipulation combined prior progress, inferred sunk effort, original task size, and completion progress. Keeping remaining effort equal controlled stated future costs; however, it did not isolate these other dimensions. Future designs could vary original workload and remaining workload separately, and compare described progress with work actually executed by an agent before an interruption.

Moreover, the design contained only ten template families. Setting variants and repeated requests increased observation counts without producing an equally large number of independent scenarios.

DOI: 10.54254/2753-7064/2026.BA39267

Both clustered regressions and family-bootstrap intervals may be fragile with so few clusters. A broader scenario set, planned before examining outcomes, would strengthen assessment of generalization. The categorical analysis was added after inspection and must remain explicitly exploratory.

Furthermore, response exclusions and implementation choices constrain comparison. Invalid rates varied from 2.9% to 9.1% by model, and the failures included truncation, empty answers, and nonconforming text. Valid-only estimates may be selective, even though extreme bounds preserved the largest descriptive patterns. Nemotron's thinking configuration also differed from the unmodified defaults of the other endpoints. This study thus cannot separate model identity from provider behavior or reasoning configuration.

Finally, endpoint aliases may change over time, and the experiment sampled a single collection period. The observed A preference reinforces the need for label and position controls, consistent with prior research on option-order sensitivity [4]. Future work could test alternative labels, require selection by objective name, compare several prompt phrasings, and repeat collection under independently frozen endpoint configurations. Such extensions should distinguish genuine changes in choice from changes in format compliance.

6. Conclusion

In 3,000 planned choices with equal stated remaining costs and rewards, described prior progress was associated with sharply different response patterns across three model endpoints. Muse mainly avoided partially completed targets, Kimi largely favored them, and Nemotron showed a smaller increase at 75% progress. The main descriptive contrasts were preserved across template-family checks and under extreme assignments of invalid outcomes. However, control-condition label preference, reasoning-setting differences, missing outcomes, and limited template diversity restrict interpretation. These results support a bounded conclusion about prompted objective selection; they do not establish human-like motivation, memory, or actual task resumption.

References

[1] Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning, 202*, 337–371. https://proceedings.mlr.press/v202/aher23a.html

[2] Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences, 122*(24), e2501660122. https://doi.org/10.1073/pnas.2501660122

[3] Ghibellini, R., & Meier, B. (2025). Interruption, recall and resumption: A meta-analysis of the Zeigarnik and Ovsiankina effects. *Humanities and Social Sciences Communications, 12*, 962.

[4] Pezeshkpour, P., & Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. *Findings of the Association for Computational Linguistics: NAACL 2024*, 2006–2017.

[5] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351.

[6] Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120*(6), e2218523120.

[7] Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2024). Large language models are not robust multiple choice selectors. *International Conference on Learning Representations*.

[8] Statsmodels developers. (n.d.). statsmodels.genmod.generalized_linear_model.GLM.fit. *Statsmodels 0.15.0 documentation*. https://www.statsmodels.org/stable/generated/statsmodels.genmod.generalized_linear_model.GLM.fit.html

DOI: 10.54254/2753-7064/2026.BA39267

[9] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://www.jstor.org/stable/4615733
[10] OpenAI. (n.d.). Codex [AI coding assistant]. https://openai.com/codex/