Post-training makes large language models less human-like

Marcel Binz[1*], Elif Akata[1], Abdullah Almaatouq[2], Mohammed Alsobay[2], Oleksii Ariasov[3], Franziska Brändle[4], David Broska[5], Jason W. Burton[6,7], Nuno Busch[8], Frederick Callaway[9], Vanessa Cheung[10], Brian Christian[4], Julian Coda-Forno[1,8], Can Demircan[1], Vittoria Dentella[11], Maria K. Eckstein[12], Noémi Éltető[13], Michael Franke[3], Thomas L. Griffiths[14], Fritz Günther[15], Susanne Haridi[16], Sebastian Hellmann[8], Stefan Herytash[16], Linus Hof[8], Eleanor Holton[4], Isabelle Hoxha[17,18], Zak Hussain[19], Akshay Jagadish[1], Elif Kara[16], Valentin Kriegmair[7], Evelina Leivada[20,21], Li Ji-An[22], Tobias Ludwig[3], Maximilian Maier[10,25], Marcelo G. Mattar[9], Marvin Mathony[1], Alireza Modirshanechi[1,13], Robin Na[2], Mariia Nadverniuk[3], Antonios Nasioulas[17,23], Surabhi S. Nath[13], Helen Niemeyer[24], Kate Nussenbaum[37], Sebastian Olschewski[19,25], Thorsten Pachur[8], Stefano Palminteri[17], Aliona Petrenco[15], Camille V. Phaneuf-Hadd[26], Angelo Pirrone[27], Manuel Rausch[28,29,30], Laura Raveling[15], Shashank Reddy[4,31], Milena Rmus[1], Evan M. Russek[14], Tankred Saanum[13,1], Kai Sandbrink[4], Louis Schiekiera[15,24], Johannes A. Schubert[1,4], Luca M. Schulze Buschoff[1], Nishad Singhi[32], Leah H. Somerville[26], Mikhail S. Spektor[33,25], Xin Sui[13,3], Christopher Summerfield[4], Mirko Thalmann[1], Anna I. Thoma[7], Taisiia Tikhomirova[7], Vuong Truong[34], Polina Tsvilodub[3],

Konstantinos Voudouris[1], Kristin Witte[1], Shuchen Wu[36],
Dirk U. Wulff[7,19], Hua-Dong Xiong[35], Songlin Xu[22], Lance Ying[2], Xinyu Zhang[22],
Jian-Qiao Zhu[14,38], Eric Schulz[1]

[1]Helmholtz Munich, [2]Massachusetts Institute of Technology, [3]University of Tübingen,
[4]University of Oxford, [5]Stanford University, [6]University of Copenhagen,
[7]Max Planck Institute for Human Development, [8]Technical University of Munich, [9]New York University,
[10]University College London, [11]University of Pavia, [12]Google DeepMind,
[13]Max Planck Institute for Biological Cybernetics, [14]Princeton University, [15]Humboldt-Universität zu Berlin,
[16]LMU Munich, [17]École normale supérieure, [18]Leiden University, [19]University of Basel,
[20]Autonomous University of Barcelona, [21]Institució Catalana de Recerca i Estudis Avançats (ICREA),
[22]University of California San Diego, [23]Paris School of Economics, [24]Freie Universität Berlin,
[25]University of Warwick, [26]Harvard University, [27]University of Liverpool, [28]Hochschule Rhein-Waal,
[29]Katholische Universität Eichstätt-Ingolstadt, [30]Alpe-Adria-Universität Klagenfurt,
[31]Singapore Management University, [32]TU Darmstadt, [33]VinUniversity, [34]Taipei Medical University,
[35]Georgia Institute of Technology, [36]Allen Institute, [37]Boston University, [38]The University of Hong Kong,
[39]Hunter College, City University of New York.

Large language models (LLMs) are increasingly used as surrogates for human participants, but it remains unclear which models best capture human behavior and why. To address this, we introduce PSYCH-201, a novel dataset that enables us to measure behavioral alignment at scale. We find that post-training – the stage that turns base models into useful assistants – consistently reduces alignment with human behavior across model families, sizes, and objectives. Moreover, this misalignment widens in newer model generations even as base models continue to improve. Finally, we find that persona-induction – a popular technique for eliciting human-like behavior by conditioning models on participant-specific information – does not improve predictions at the level of individuals. Taken together, our results suggest that the very processes that are currently employed to turn LLMs into useful assistants also make them less accurate models of human behavior.

Large language models (LLMs) such as ChatGPT, Claude, and Gemini have rapidly transformed the landscape of science and society, serving as powerful tools for writing, coding, and reasoning (*1, 2*). Most current development is geared toward turning these models into useful assistants that provide normatively correct responses. Yet one of their most far-reaching applications lies elsewhere: faithfully mimicking human behavior, including its errors, variance, and the factors that shape it (*3–7*). These human-like models could be applied to simulate patient responses in mental health care, which, for instance, could train psychiatrists on challenging clinical cases in-silico. They could make it possible to anticipate how individuals and populations will respond to policy interventions even before those interventions are deployed. They could help model student learning trajectories in educational settings, thereby guiding the design of more effective and personalized curricula.

However, the extent to which LLMs actually resemble human behavior remains disputed. While some studies report strong behavioral alignment between model outputs and human responses (*8–11*), others find more mixed results or even major divergences (*12–15*).[1] This raises fundamental questions: Which model properties drive alignment with human behavior? In which domains do

[1]Throughout this article, we use the term behavioral alignment to refer to the behavioral similarity between models and humans, which is distinct from (but related to) the broader usage of alignment in safety research, where it refers to model compliance with human preferences and values.

they accurately capture human responses? Can they adapt to the characteristics of individual people?

To answer these questions, we need a testbed that covers a broad spectrum of behaviors. The present paper introduces such a resource in the form of PSYCH-201, a large-scale dataset of natural language transcripts from behavioral experiments. PSYCH-201 was collected through an open research collaboration and crowdsourced effort, resulting in a dataset 3.5× larger than its predecessor PSYCH-101, a more diverse participant population, and broader coverage of experimental paradigms.

Equipped with this, we then evaluated a wide range of LLMs on their ability to predict human responses in PSYCH-201. These models come in different shapes and forms. Base models are the output of pretraining, in which the model learns to predict the next word in large text corpora. Post-trained models take a base model and further adapt it toward more specific objectives, such as instruction-tuning (teaching models to follow user requests), reasoning (training models to produce normatively correct responses alongside step-by-step reasoning traces), or vision (extending models to process images in addition to text). We find that:

1. Post-training consistently reduces human-likeness. This effect holds across model families and applies to all post-training objectives, including instruction-tuning, reasoning, and vision.

2. Base models continue to improve across generations, i.e. newer models are generally more aligned.

3. Post-training misalignment – defined as the alignment difference between a base model and its post-trained counterpart – widens in newer models.

4. The largest post-training misalignment occurs in the domains of psycholinguistics and reasoning.

5. Persona-induction, a popular technique for eliciting more human-like behavior by conditioning models on participant-specific information, does not improve predictions at the level of individuals.

Taken together, our findings have important implications for using LLMs as behavioral surrogates. Most prior studies rely on post-trained models, given their accessibility, user-friendliness, and ease of prompting. Yet our results suggest that this practice is suboptimal if the goal is to produce

systems that are similar to humans, and they motivate new approaches to post-training that preserve the behavioral alignment found in base models. Looking beyond the presented results, PSYCH-201 offers a useful resource for meta-analyses, automated scientific discovery, and the development of large-scale cognitive models.

PSYCH-201

To systematically evaluate the behavioral alignment of LLMs with human responses, we introduce PSYCH-201, a novel dataset consisting of natural language transcripts from behavioral experiments. PSYCH-201 contains trial-by-trial data from individual participants, with each data sequence corresponding to the transcript of an entire experimental session (including instructions, stimuli, responses, and any task-relevant context presented during the session; see Fig. 1a for an example).

PSYCH-201 was collected through an open research collaboration, resulting in a dataset of unprecedented scale and diversity. It includes data from 208, 021 participants, 25, 906, 599 behavioral responses, and hundreds of experiments, making it 3.5× larger than its predecessor PSYCH-101 and 13× larger than the average mega-study (*16*); see Fig. 1b. It is furthermore more diverse than PSYCH-101, both in terms of experimental paradigms (see Fig. 1c) and participant demographics. Notably, it incorporates several developmental and cross-cultural studies, thereby increasing demographic coverage (see Fig. 1d-e). Each data point is annotated with detailed meta-data, including age, nationality, questionnaire responses, and other markers.

Post-training makes LLMs less human-like

We evaluated models from three major families on PSYCH-201: Qwen3, a state-of-the-art open-source model family (*18*); Llama3.X, a model family with a broad ecosystem (*19*); and Olmo3.X, a fully open and reproducible model family (*20*). For each model, we measured behavioral alignment using the negative log-likelihood (NLL) of individual human responses under predictions of the respective model (with lower values indicating closer alignment with human behavior). To quantify the effect of post-training, we report effect sizes (Cohen's d) relative to the corresponding base model, averaged across all experiments unless otherwise noted. Likewise, we measure the benefit of meta-data through effect sizes between models prompted with and without participant-specific

(a)

In the following study, you will see complex words such as *airport* or *warzone* in the browser window. You will have to judge for each word whether it exists as an English word or not.

You will see these words one after another. If the word presented to you exists, press the **E** key on your keyboard. If the word does not exist, press the **R** key.

`descentmajor`	You press <<R>>.
`bunkpunk`	You press <<R>>.
`seaplane`	You press <<R>>.
`bloodshed`	You press <<E>>.
`westscheme`	You press <<R>>.

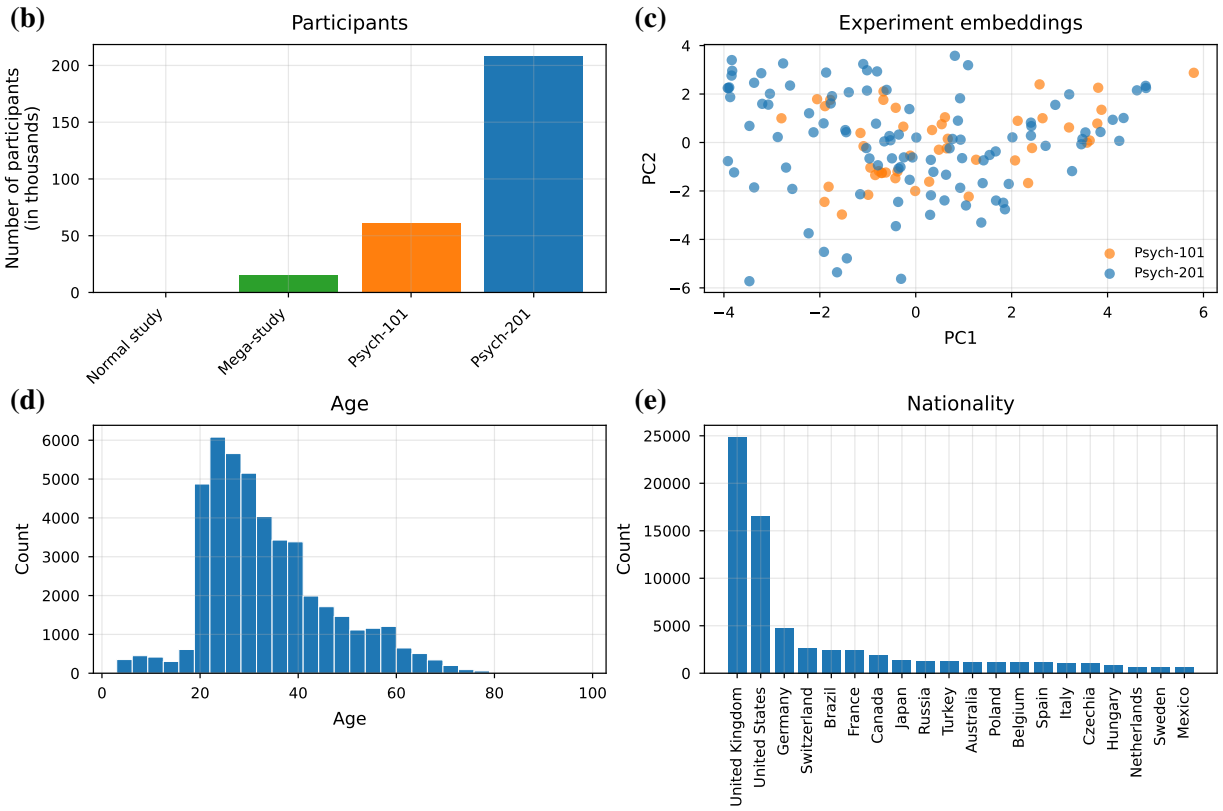

Figure 1: Psych-201 data example and statistics. (a) Example of a single data point, i.e. a transcript of one experimental session. (b) Number of participants across datasets, showing that Psych-201 reaches a scale multiple times larger than typical behavioral datasets (*16*). (c) BERT (*17*) embeddings for experiments from Psych-101 (orange) and Psych-201 (blue). Embeddings were projected onto two dimensions using principal component analysis. Psych-201 covers a substantially broader spectrum of experimental paradigms than its predecessor. (d) Histogram over age in Psych-201. (e) Histogram over nationality in Psych-201.

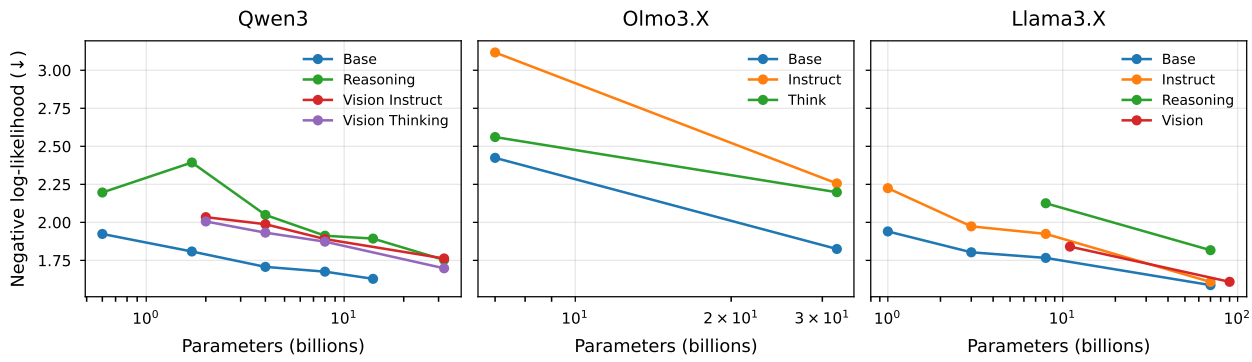

Figure 2: Behavioral alignment for different post-training objectives. Negative log-likelihood of human responses from Psych-201 for models from the Qwen3, Olmo3.X, and Llama3.X families (lower means more aligned). Post-trained models (shown in non-blue) exhibit consistently lower alignment with human responses across all objectives, families, and sizes than their base model counterpart (shown in blue).

meta-data.

We observed a consistent and robust effect across model families and sizes: post-training reduces behavioral alignment (see Fig. 2). This effect holds across all major post-training objectives. In the aggregate, instruction-tuned models (average $d = 0.11$), reasoning models (average $d = 0.14$), and vision models (average $d = 0.07$) are all less aligned than their corresponding base models. Furthermore, the base model outperforms its post-trained counterpart in almost every direct comparison. The most aligned model overall was Qwen2.5-72B (average NLL = 1.557; note that base models are not available in larger sizes for the latest Qwen generations). We present additional analysis in Fig. S1, showing this post-training effect is robust across different prompt formats.

One potential explanation for this effect is that post-trained models simply produce more deterministic outputs, thereby failing to capture the noisiness of human behavior (*21*). If this were the case, however, we would expect post-trained models to exhibit equal or higher accuracy since the (unchanged) mode of the output distribution would still align with human responses. Further analyses presented in Fig. S2 indicate that this is not the case, thereby ruling out increased determinism as a sole driver of the observed misalignment.

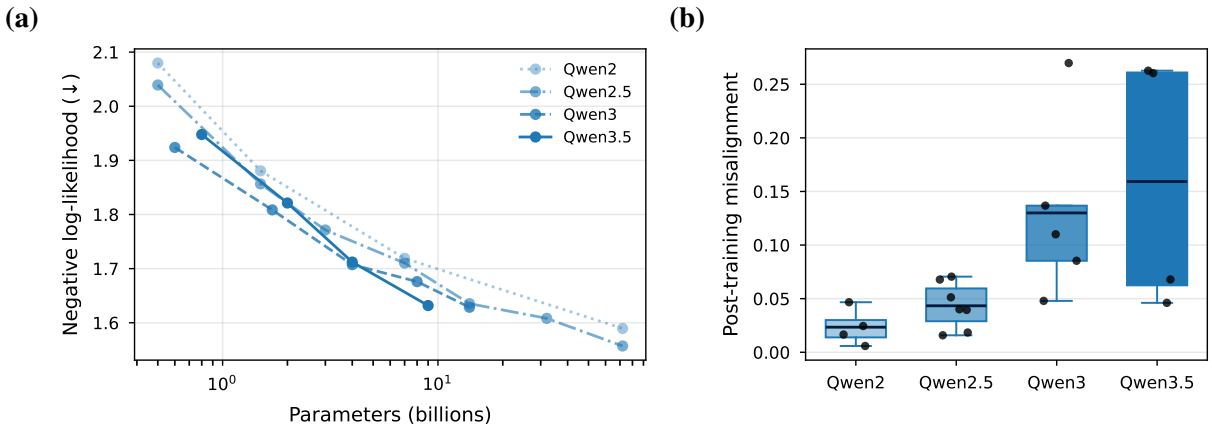

Figure 3: Behavioral alignment across Qwen generations. (a) Behavioral alignment for base models improves across generations. (b) The post-training misalignment between instruction-tuned models and their corresponding base model increases across generations.

Behavioral alignment across model generations

How does behavioral alignment change with newer model generations? This is especially interesting as in some domains there is evidence that alignment with human behavior plateaus – or even declines – as models become more powerful (*22–25*). We find that the base models, in fact, continue to improve in their behavioral alignment across model generations (see Fig. 3a for models from the Qwen family). However, we also find that the post-training misalignment gap widens in newer model generations (see Fig. 3b). For instance, the gap for instruction-tuned models is relatively modest in Qwen2 (average $d = 0.02$) and Qwen2.5 (average $d = 0.04$), but increases in Qwen3 (average $d = 0.13$) and further in Qwen3.5 (average $d = 0.16$), suggesting that ongoing developments in post-training amplify the divergence from human-like behavior.

Mapping behavioral alignment across experimental domains

Is there any structure in which types of experimental characteristics lead to more or less behavioral alignment? To investigate this, we grouped the experiments in PSYCH-201 by domain and analyzed the post-training misalignment. We find that post-training misalignment is present across all domains, indicating that it is a domain-general effect (see Fig. 4). However, the magnitude of misalignment varies across domains and models.

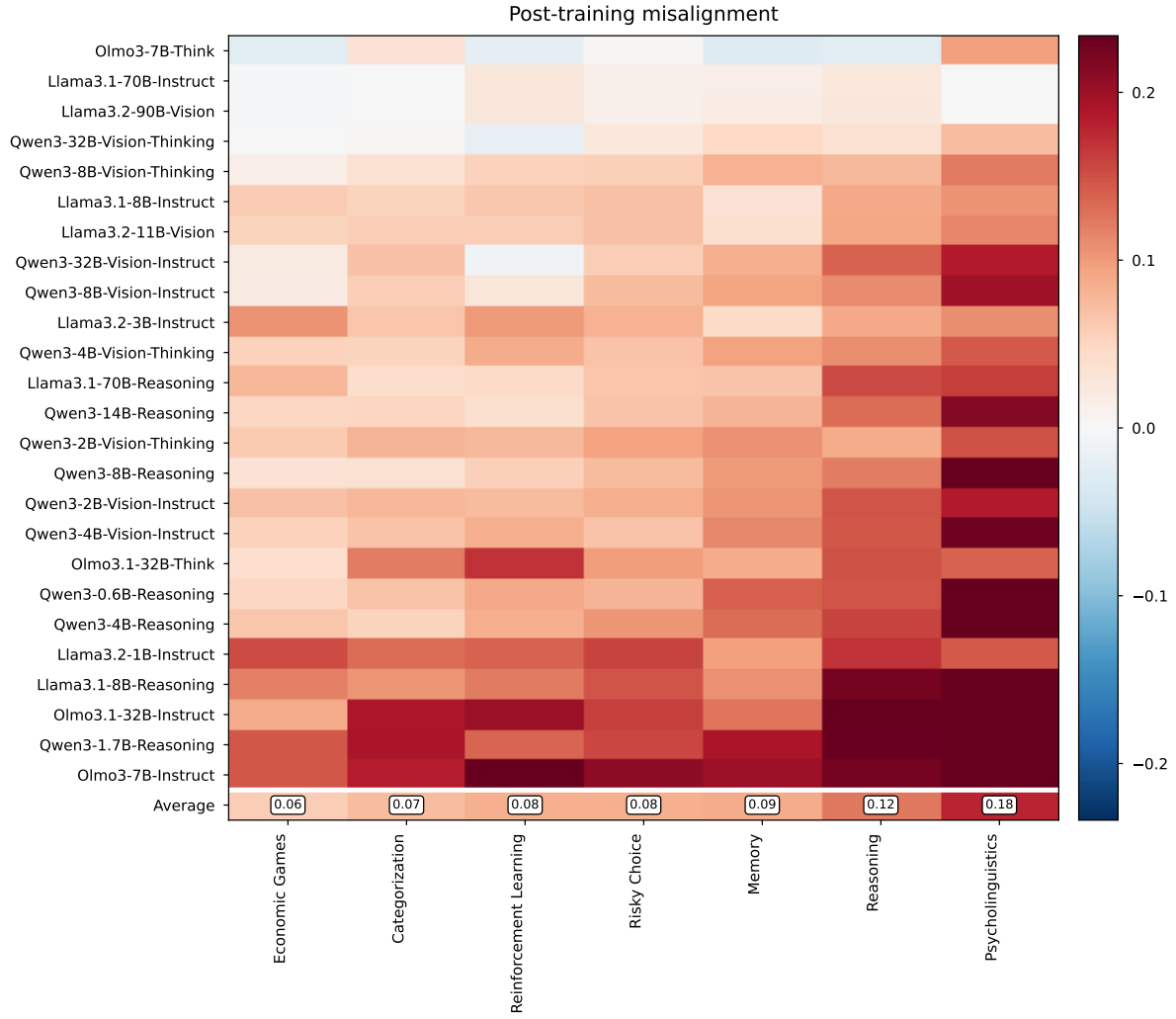

Figure 4: Post-training misalignment by experiment domain. Each individual tile shows average effect sizes (grouped by domain) for a given model. Positive values indicate a negative effect of post-training.

The domains with the highest post-training misalignment are reasoning and psycholinguistics. A natural explanation for this is that base models and post-trained models are optimized for different objectives: base models are ultimately models of human language, and should therefore be especially well aligned with human behavior on psycholinguistic tasks. Post-training techniques, such as reinforcement learning from human feedback, on the other hand, are designed to maximize user engagement, thereby shifting models away from their original objective. The same may hold for reasoning tasks: human decision-making is shaped by heuristics and biases (*26, 27*), which might be captured by base models but are then overwritten by reasoning post-training, which optimizes for normatively correct responses. Thus, the very processes that are currently employed to turn these models into useful assistants may also make them less accurate models of human behavior.

No benefit of persona-induction

Persona-induction, i.e. conditioning a model on information about a particular individual, has become a popular approach for eliciting more human-like behavior from LLMs (*28–31*). While previous studies have shown that persona-induction can produce human-aligned responses at the population level (*3, 4, 32*), it has not been systematically evaluated whether it actually improves the prediction of responses at the level of individuals. Psych-201's rich participant-level meta-data enables the first systematic test of persona-induction at scale.

Following prior work, we adopted an interview-style prompting format (*33*) in which each experimental transcript was preceded by question-answer pairs describing the participant (e.g. "Interviewer: What is your age? Interviewee: 35."). We included information about age, gender, nationality, education, clinical diagnosis, and questionnaire statistics where available.

In general, we found only a minor effect of persona-induction. The meta-data benefit ranged from $d = -0.02$ (Llama3.2-3B-Instruct) to $d = 0.02$ (Llama3.1-70B). This pattern held for both base and instruction-tuned models (see Fig. 5), and persisted when restricting the analysis to developmental experiments, where age-related differences in behavior are known to be present and informative. Taken together, these findings challenge the validity of persona-induction techniques, at least for modeling individual-level behavior. While such methods may change model outputs in ways that appear more human-like on the population level, they do not necessarily help to capture

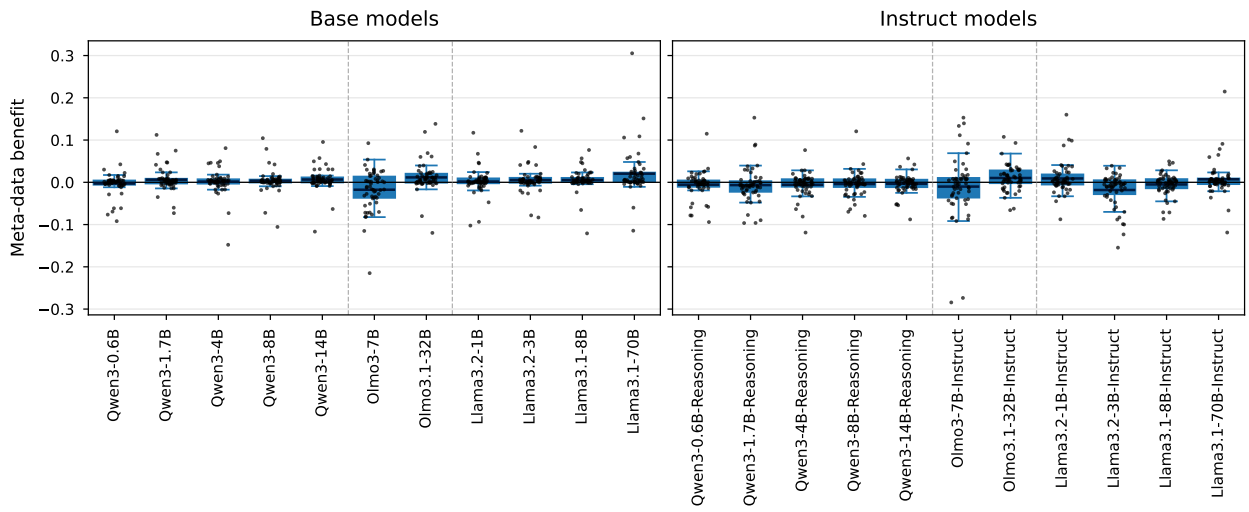

Figure 5: Effect of persona-induction on individual-level behavioral prediction. Each subplot shows the effect of adding participant-specific meta-data to the prompt for a given model. Positive values indicate that persona-induction improves prediction, whereas negative values indicate that it worsens prediction.

the behavior of individual people.

Discussion

In this study, we have systematically evaluated the alignment between human behavior and LLMs. The results show that post-training consistently reduces behavioral alignment across model families, sizes, and post-training objectives. While base models still continue to improve across model generations, post-training misalignment actually increases in newer models, highlighting the pressing nature of this issue. These results were enabled by PSYCH-201, a dataset assembled through an open research collaboration that brought together contributions from across the behavioral sciences, yielding a resource of unprecedented scale.

The observation that post-trained models exhibit reduced behavioral alignment with humans both connects to and extends several lines of earlier work. First, previous work has shown that early language models exhibited human-like cognitive biases, but that these patterns tend to disappear – and were instead replaced with more rational behaviors – in newer models that have undergone more extensive post-training (*13, 14, 34, 35*). Second, instruction-tuned models have been found

to produce less human-like text than their base counterparts (*36*). We extend this finding from text generation to behavioral responses, and demonstrate that it holds across a broad range of post-training objectives. Lastly, looking beyond behavioral alignment, a similar observation has also been made in the psycholinguistic literature, where base models better predict human reading times than their instruction-tuned counterparts (*37*).

More broadly, our results can be viewed as a form of the alignment tax (*38*) – a phenomenon whereby post-training can degrade model capabilities acquired during pretraining. While recent work has proposed methods to mitigate such issues on common benchmarks (*39*), our findings suggest that these mitigations do not extend to behavioral alignment with humans. It might thus be asked whether the observed misalignment reflects a fundamental limitation of post-training or merely shortcomings of current techniques. To probe this question, we evaluated Centaur, a model fine-tuned on a subset of tasks from Psych-201 (i.e. those previously included in the Psych-101 dataset). Centaur exhibited a consistent increase in behavioral alignment on held-out, novel tasks that were not part of its training set (average $d = 0.28$, SEM = 0.10), suggesting a substantial degree of generalization and demonstrating that post-training can, in fact, help.

These findings point to an important direction for future work: developing post-training methods that preserve the behavioral fidelity of base models while retaining the practical benefits of post-trained models. This includes not only capturing general patterns of human behavior, but also the stochasticity and individual-level variability that characterize human responses. The latter is especially critical when simulating behavior at the level of specific individuals or subpopulations. Psych-201 provides rich participant-level metadata, which in principle allows us to build models that accomplish exactly this.

There is immense potential – for both scientific and applied reasons – for LLMs that closely mimic human behavior. If we had models that accurately simulate human learners, we could use them to design optimal educational curricula. If we had models that emulate human responses in mental health contexts, we could use them as training tools for psychiatrists. If we had models that accurately model public responses to policy changes, we could support more informed governance. To realize this potential, however, we need to move beyond the current paradigm in which post-training is exclusively optimized for normative correctness. Psych-201 provides the foundation for evaluating behavioral alignment at scale, thereby directly paving the way towards this goal.

Funding

This work was supported by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@FZJ partition. This work was funded by the German Research Foundation (DFG), Emmy-Noether grant "What's in a name?" (project number 459717703), awarded to Fritz Günther. This work was funded by the German Research Foundation (DFG), research grant "A computational implementation of the Swinging Lexical Network model of language production" (project number 532390335), awarded to Fritz Günther. Evelina Leivada acknowledges funding by the Ministerio de Ciencia, Innovación y Universidades (Spain) under grant agreement CNS2023-144415. This work was made possible with the support of the NOMIS Foundation (Brian Christian, Jian-Qiao Zhu, Evan M. Russek, and Thomas L. Griffiths).

References and Notes

1. R. Bommasani, *et al.*, On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

2. M. Binz, *et al.*, How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences* **122** (5), e2401227121 (2025).

3. J. S. Park, *et al.*, Generative agents: Interactive simulacra of human behavior, in *Proceedings of the 36th annual acm symposium on user interface software and technology* (2023), pp. 1–22.

4. J. S. Park, *et al.*, Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109* (2024).

5. M. Binz, *et al.*, A foundation model to predict and capture human cognition. *Nature* **644** (8078), 1002–1009 (2025).

6. Z. Cui, N. Li, H. Zhou, A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science* **5** (8), 627–634 (2025).

7. J. Hullman, D. Broska, H. Sun, A. Shaw, This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. *arXiv preprint arXiv:2602.15785* (2026).

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

9. R. Marjieh, I. Sucholutsky, P. van Rijn, N. Jacoby, T. L. Griffiths, Large language models predict human sensory judgments across six modalities. *Scientific Reports* **14** (1), 21445 (2024).

10. J. Hu, K. Mahowald, G. Lupyan, A. Ivanova, R. Levy, Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences* **121** (36), e2400917121 (2024).

11. A. K. Lampinen, *et al.*, Language models, like humans, show content effects on reasoning tasks. *PNAS nexus* **3** (7), pgae233 (2024).

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

13. T. Hagendorff, S. Fabi, M. Kosinski, Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* **3** (10), 833–838 (2023).

14. Y. Chen, T. X. Liu, Y. Shan, S. Zhong, The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences* **120** (51), e2316205120 (2023).

15. Y. Gao, D. Lee, G. Burtch, S. Fazelpour, Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences* **122** (24), e2501660122 (2025).

16. L. Kastrati, *et al.*, Agreement between mega-trials and smaller trials: a systematic review and meta-research analysis. *JAMA network open* **7** (9), e2432296 (2024).

17. B. Warner, *et al.*, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2025), pp. 2526–2547.

18. A. Yang, *et al.*, Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

19. A. Grattafiori, *et al.*, The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

20. T. Olmo, *et al.*, Olmo 3. *arXiv preprint arXiv:2512.13961* (2025).

21. R. Kirk, *et al.*, Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452* (2023).

22. D. Linsley, *et al.*, Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *Advances in Neural Information Processing Systems* **36**, 28873–28891 (2023).

23. B.-D. Oh, W. Schuler, Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics* **11**, 336–350 (2023).

24. A. J. Wu, R. Liu, X. Bai, T. L. Griffiths, Large Language Models Develop Novel Social Biases Through Adaptive Exploration. *arXiv preprint arXiv:2511.06148* (2025).

25. K. M. Collins, *et al.*, Evaluating Language Models' Evaluations of Games. *arXiv preprint arXiv:2510.10930* (2025).

26. G. Gigerenzer, W. Gaissmaier, Heuristic decision making. *Annual review of psychology* **62** (2011), 451–482 (2011).

27. M. Binz, S. J. Gershman, E. Schulz, D. Endres, Heuristics from bounded meta-learned inference. *Psychological review* **129** (5), 1042 (2022).

28. L. Salewski, S. Alaniz, I. Rio-Torto, E. Schulz, Z. Akata, In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems* **36**, 72044–72057 (2023).

29. S. Wu, *et al.*, HumanLM: Simulating Users with State Alignment Beats Response Imitation. *arXiv preprint arXiv:2603.03303* (2026).

30. D. Paglieri, L. Cross, W. A. Cunningham, J. Z. Leibo, A. S. Vezhnevets, Persona Generators: Generating Diverse Synthetic Personas at Scale. *arXiv preprint arXiv:2602.03545* (2026).

31. Y. Ma, *et al.*, Synthetic Interaction Data for Scalable Personalization in Large Language Models. *arXiv preprint arXiv:2602.12394* (2026).

32. L. P. Argyle, *et al.*, Out of one, many: Using language models to simulate human samples. *Political Analysis* **31** (3), 337–351 (2023).

33. M. Lutz, I. Sen, G. Ahnert, E. Rogers, M. Strohmaier, The prompt makes the person (a): A systematic evaluation of sociodemographic persona prompting for large language models. *arXiv preprint arXiv:2507.16076* (2025).

34. V. Cheung, M. Maier, F. Lieder, Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences* **122** (25), e2412015122 (2025).

35. E. Shapira, M. Tennenholtz, R. Reichart, Alignment Makes Language Models Normative, Not Descriptive. *arXiv preprint arXiv:2603.17218* (2026).

36. A. Reinhart, *et al.*, Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences* **122** (8), e2422455122 (2025).

37. T. Kuribayashi, Y. Oseki, T. Baldwin, Psychometric predictive power of large language models, in *Findings of the Association for Computational Linguistics: NAACL 2024* (2024), pp. 1983–2005.

38. L. Ouyang, *et al.*, Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35**, 27730–27744 (2022).

39. Y. Lin, *et al.*, Mitigating the alignment tax of rlhf, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (2024), pp. 580–606.

40. P. Aggarwal, E. A. Cranford, M. Tambe, C. Lebiere, C. Gonzalez, Deceptive signaling: Understanding human behavior against signaling algorithms, in *Cyber deception: Techniques, strategies, and human aspects* (Springer), pp. 83–95 (2022).

41. M. Agrawal, J. C. Peterson, J. D. Cohen, T. L. Griffiths, Stress, intertemporal choice, and mitigation behavior during the COVID-19 pandemic. *Journal of Experimental Psychology: General* **152** (9), 2695 (2023).

42. E. Akata, *et al.*, Playing repeated games with large language models. *Nature Human Behaviour* **9** (7), 1380–1390 (2025).

43. M. Alsobay, D. G. Rand, D. J. Watts, A. Almaatouq, Integrative experiments identify how punishment affects welfare in public goods games. *Science* **392** (6794), eaeb5280 (2026).

44. H. Anlló, *et al.*, Comparing experience-and description-based economic preferences across 11 countries. *Nature Human Behaviour* **8** (8), 1554–1567 (2024).

45. F. Anvari, S. Billinger, P. P. Analytis, V. R. Franco, D. Marchiori, Testing the convergent validity, domain generality, and temporal stability of selected measures of people's tendency to explore. *Nature Communications* **15** (1), 7721 (2024).

46. E. Awad, *et al.*, The moral machine experiment. *Nature* **563** (7729), 59–64 (2018).

47. J. M. van Baar, M. R. Nassar, W. Deng, O. FeldmanHall, Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour* **6** (3), 404–414 (2022).

48. J. M. Barnby, N. Raihani, P. Dayan, Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. *Cognition* **225**, 105098 (2022).

49. S. Bavard, M. Lebreton, M. Khamassi, G. Coricelli, S. Palminteri, Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nature communications* **9** (1), 4503 (2018).

50. S. Bavard, A. Rustichini, S. Palminteri, Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. *Science Advances* **7** (14), eabe0340 (2021).

51. S. Bavard, S. Palminteri, The functional form of value normalization in human reinforcement learning. *Elife* **12**, e83891 (2023).

52. S. Bhatia, Exploring variability in risk taking with large language models. *Journal of Experimental Psychology: General* **153** (7), 1838 (2024).

53. F. Brändle, L. J. Stocks, J. B. Tenenbaum, S. J. Gershman, E. Schulz, Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour* **7** (9), 1481–1489 (2023).

54. A. D. Breslav, *et al.*, Shuffle the decks: Children are sensitive to incidental nonrandom structure in a sequential-choice task. *Psychological Science* **33** (4), 550–562 (2022).

55. J. W. Burton, A. J. Harris, P. Shah, U. Hahn, Optimism where there is none: Asymmetric belief updating observed with valence-neutral life events. *Cognition* **218**, 104939 (2022).

56. N. Busch, T. Geyer, A. Zinchenko, Individual peak alpha frequency does not index individual differences in inhibitory cognitive control. *Psychophysiology* **61** (8), e14586 (2024).

57. P. Castro-Rodrigues, *et al.*, Explicit knowledge of task structure is a primary determinant of human model-based action. *Nature human behaviour* **6** (8), 1126–1141 (2022).

58. V. Chambon, *et al.*, Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour* **4** (10), 1067–1079 (2020).

59. S. Ciranka, W. van den Bos, Internal uncertainty impacts social information use in risky choice across adolescence. *Communications Psychology* **3** (1), 137 (2025).

60. A. O. Cohen, K. Nussenbaum, H. M. Dorfman, S. J. Gershman, C. A. Hartley, The rational use of causal inference to guide reinforcement learning strengthens with age. *npj Science of Learning* **5** (1), 16 (2020).

61. J. H. Decker, A. R. Otto, N. D. Daw, C. A. Hartley, From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological science* **27** (6), 848–858 (2016).

62. C. Demircan, *et al.*, Evaluating alignment between humans and neural network representations in image-based learning tasks. *Advances in Neural Information Processing Systems* **37**, 122406–122433 (2024).

63. A. Dezfouli, K. Griffiths, F. Ramos, P. Dayan, B. W. Balleine, Models that learn how humans learn: The case of decision-making and its disorders. *PLoS computational biology* **15** (6), e1006903 (2019).

64. M. Dubois, T. U. Hauser, Value-free random exploration is linked to impulsivity. *Nature Communications* **13** (1), 4542 (2022).

65. I. Evangelidis, J. Levav, I. Simonson, The upscaling effect: How the decision context influences tradeoffs between desirability and feasibility. *Journal of Consumer Research* **50** (3), 492–509 (2023).

66. H. Fan, S. J. Gershman, E. A. Phelps, Trait somatic anxiety is associated with reduced directed exploration and underestimation of uncertainty. *Nature Human Behaviour* **7** (1), 102–113 (2023).

67. C. Feher da Silva, T. A. Hare, Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour* **4** (10), 1053–1066 (2020).

68. M. Franke, P. Tsvilodub, F. Carcassi, Bayesian statistical modeling with predictors from LLMs. *arXiv preprint arXiv:2406.09012* (2024).

69. M. Franke, J. Degen, Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one* **11** (5), e0154854 (2016).

70. R. Frey, A. Pedroni, R. Mata, J. Rieskamp, R. Hertwig, Risk preference shares the psychometric structure of major psychological traits. *Science advances* **3** (10), e1701381 (2017).

71. C. M. Gillan, M. Kosinski, R. Whelan, E. A. Phelps, N. D. Daw, Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *elife* **5**, e11305 (2016).

72. J. Groß, B. K. Kreis, H. Blank, T. Pachur, Knowledge updating in real-world estimation: Connecting hindsight bias and seeding effects. *Journal of Experimental Psychology: General* **152** (11), 3167 (2023).

73. F. Günther, M. Marelli, Trying to make it work: Compositional effects in the processing of compound "nonwords". *Quarterly Journal of Experimental Psychology* **73** (7), 1082–1091 (2020).

74. F. Günther, M. A. Petilli, M. Marelli, Semantic transparency is not invisibility: A computational model of perceptually-grounded conceptual combination in word processing. *Journal of Memory and Language* **112**, 104104 (2020).

75. F. Günther, M. Marelli, Patterns in CAOSS: Distributed representations predict variation in relational interpretations for familiar and novel compound words. *Cognitive Psychology* **134**, 101471 (2022).

76. F. Guenther, M. Marelli, S. Tureski, M. A. Petilli, ViSpa (Vision Spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psychological Review* **130** (4), 896 (2023).

77. V. Dentella, F. Günther, E. Leivada, Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences* **120** (51), e2309583120 (2023).

78. V. Dentella, F. Günther, E. Murphy, G. Marcus, E. Leivada, Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports* **14** (1), 28083 (2024).

79. V. Pugacheva, F. Günther, Lexical choice and word formation in a taboo game paradigm. *Journal of Memory and Language* **135**, 104477 (2024).

80. M. P. Gunadi, I. Evangelidis, The impact of historical price information on purchase deferral. *Journal of Marketing Research* **59** (3), 623–640 (2022).

81. N. Haines, *et al.*, Anxiety modulates preference for immediate rewards among trait-impulsive individuals: A hierarchical Bayesian analysis. *Clinical Psychological Science* **8** (6), 1017–1036 (2020).

82. S. Haridi, E. Schulz, M. Thalmann, Context Size and Set Size Effects: The Relevance of Specific Cues When Searching Long-Term Memory. *Computational Brain & Behavior* **9** (1), 1–33 (2026).

83. K. Nussenbaum, P. L. Katzman, H. Lu, S. Zorowitz, C. A. Hartley, Sensitivity to the instrumental value of choice increases across development. *Psychological Science* **35** (8), 933–947 (2024).

84. J. Heffner, O. FeldmanHall, A probabilistic map of emotional experiences during competitive social interactions. *Nature communications* **13** (1), 1718 (2022).

85. E. Holton, *et al.*, Goal commitment is supported by vmPFC through selective attention. *Nature Human Behaviour* **8** (7), 1351–1365 (2024).

86. J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, E. Gibson, A fine-grained comparison of pragmatic language understanding in humans and language models, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2023), pp. 4194–4213.

87. L. E. Hunter, E. A. Meer, C. M. Gillan, M. Hsu, N. D. Daw, Increased and biased deliberation in social anxiety. *Nature Human Behaviour* **6** (1), 146–154 (2022).

88. Z. Hussain, R. Mata, D. U. Wulff, Novel embeddings improve the prediction of risk perception. *EPJ data science* **13** (1), 38 (2024).

89. A. K. Jagadish, M. Binz, T. Saanum, J. X. Wang, E. Schulz, Zero-shot compositional reinforcement learning in humans. *Trials* **1**, 5 (2023).

90. R. A. Jansen, A. N. Rafferty, T. L. Griffiths, A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour* **5** (6), 756–763 (2021).

91. D. R. Little, R. M. Shiffrin, S. M. Laham, Function estimation: Quantifying individual differences of hand-drawn functions. *Memory & Cognition* **53** (1), 242–261 (2025).

92. J. A. Marshall, A. Reina, C. Hay, A. Dussutour, A. Pirrone, Magnitude-sensitive reaction times reveal non-linear time costs in multi-alternative decision-making. *PLoS computational biology* **18** (10), e1010523 (2022).

93. M. Moutoussis, *et al.*, Change, stability, and instability in the Pavlovian guidance of behaviour from adolescence to young adulthood. *PLoS computational biology* **14** (12), e1006679 (2018).

94. A. Nasioulas, E. Potier, F. Cerrotti, M. Lebreton, S. Palminteri, Feedback-induced attitudinal changes in risk preferences. *Nature Communications* (2026).

95. K. Nussenbaum, M. Scheuplein, C. V. Phaneuf, M. D. Evans, C. A. Hartley, Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychology* **6** (1) (2020).

96. K. Nussenbaum, *et al.*, Novelty and uncertainty differentially drive exploration across development. *Elife* **12**, e84260 (2023).

97. S. Olschewski, M. S. Spektor, G. Le Mens, Frequent winners explain apparent skewness preferences in experience-based decisions. *Proceedings of the National Academy of Sciences* **121** (12), e2317751121 (2024).

98. S. Olschewski, T. L. Mullett, N. Stewart, Optimal allocation of time in risky choices under opportunity costs. *Cognitive Psychology* **157**, 101716 (2025).

99. S. Palminteri, G. Lefebvre, E. J. Kilford, S.-J. Blakemore, Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology* **13** (8), e1005684 (2017).

100. C. V. Phaneuf-Hadd, I. M. Jacques, C. Insel, A. R. Otto, L. H. Somerville, Characterizing age-related change in learning the value of cognitive effort. *Journal of Experimental Psychology: General* (2025).

101. A. C. Pike, Á. A. Anet, N. Peleg, O. J. Robinson, Catastrophizing and risk-taking. *Computational Psychiatry* **7** (1), 1 (2023).

102. A. Pirrone, W. Wen, S. Li, Single-trial dynamics explain magnitude sensitive decision making. *BMC neuroscience* **19** (1), 54 (2018).

103. T. C. Potter, N. V. Bryce, C. A. Hartley, Cognitive components underpinning the development of model-based learning. *Developmental Cognitive Neuroscience* **25**, 272–280 (2017).

104. G. M. Rosenbaum, H. L. Grassie, C. A. Hartley, Valence biases in reinforcement learning shift across adolescence and modulate subsequent memory. *ELife* **11**, e64620 (2022).

105. E. M. Russek, R. Moran, Y. Liu, R. J. Dolan, Q. J. Huys, Heuristics in risky decision-making relate to preferential representation of information. *Nature Communications* **15** (1), 4269 (2024).

106. R. B. Rutledge, N. Skandali, P. Dayan, R. J. Dolan, A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences* **111** (33), 12252–12257 (2014).

107. K. J. Sandbrink, L. T. Hunt, C. Summerfield, Understanding human metacontrol and its pathologies using deep neural networks. *Proceedings of the National Academy of Sciences* **123** (9), e2510334123 (2026).

108. L. Schiekiera, H. Niemeyer, Political bias in historiography-an experimental investigation of preferences for publication as a function of political orientation. *F1000Research* **14**, 320 (2025).

109. N. Shahar, *et al.*, Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS computational biology* **15** (2), e1006803 (2019).

110. K. Singh, P. Aggarwal, P. Rajivan, C. Gonzalez, Training to detect phishing emails: Effects of the frequency of experienced phishing emails, in *Proceedings of the human factors and ergonomics society annual meeting* (SAGE Publications Sage CA: Los Angeles, CA), vol. 63 (2019), pp. 453–457.

111. M. Singh, R. Richie, S. Bhatia, Representing and predicting everyday behavior. *Computational Brain & Behavior* **5** (1), 1–21 (2022).

112. M. S. Spektor, S. Gluth, L. Fontanesi, J. Rieskamp, How similarity between choice options affects decisions from experience: The accentuation-of-differences model. *Psychological review* **126** (1), 52 (2019).

113. M. S. Spektor, D. Kellen, J. Rieskamp, K. C. Klauer, Absolute and relative stability of loss aversion across contexts. *Journal of Experimental Psychology: General* **153** (2), 454 (2024).

114. L. Sun, *et al.*, Large Language Models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity* **57**, 101870 (2025).

115. P. Suthaharan, *et al.*, Paranoia and belief updating during the COVID-19 crisis. *Nature human behaviour* **5** (9), 1190–1202 (2021).

116. M. H. Tessler, M. Franke, Not unreasonable: Carving vague dimensions with contraries and contradictions, in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 40 (2018).

117. A. I. Thoma, B. R. Newell, C. Schulze, Emerging adaptivity in probability learning: How young minds and the environment interact. *Journal of Experimental Psychology: General* (2025).

118. A. I. Thoma, C. Schulze, Do children match described probabilities? The sampling hypothesis applied to repeated risky choice. *Journal of Experimental Child Psychology* **251**, 106126 (2025).

119. P. Tsvilodub, B. van Tiel, M. Franke, The role of relevance, competence, and priors for scalar inferences. *Experiments in Linguistic Meaning* **2**, 288–298 (2023).

120. H. Vandendriessche, *et al.*, Contextual influence of reinforcement learning performance of depression: evidence for a negativity bias? *Psychological Medicine* **53** (10), 4696–4706 (2023).

121. B. van Tiel, M. Franke, U. Sauerland, Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences* **118** (9), e2005453118 (2021).

122. B. van Tiel, U. Sauerland, M. Franke, Meaning and use in the expression of estimative probability. *Open Mind* **6**, 250–263 (2022).

123. K. Witte, T. Wise, Q. Huys, E. Schulz, Exploring the unexplored: Worry as a catalyst for exploratory behavior in anxiety and depression (2024).

124. K. Witte, M. Thalmann, E. Schulz, Model-based exploration is measurable across tasks but not linked to personality and psychiatric assessments. *Scientific Reports* **15** (1), 27479 (2025).

125. H. A. Xu, A. Modirshanechi, M. P. Lehmann, W. Gerstner, M. H. Herzog, Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLOS Computational Biology* **17** (6), e1009070 (2021).

126. S. Xu, X. Zhang, Augmenting human cognition with an AI-mediated intelligent visual feedback, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–16.

127. L. Ying, *et al.*, The neuro-symbolic inverse planning engine (nipe): Modeling probabilistic social inferences from linguistic inputs. *arXiv preprint arXiv:2306.14325* (2023).

128. J.-Q. Zhu, J. C. Peterson, B. Enke, T. L. Griffiths, Capturing the complexity of human strategic decision-making with machine learning. *Nature Human Behaviour* **9** (10), 2114–2120 (2025).

129. O. Zika, K. Wiech, A. Reinecke, M. Browning, N. W. Schuck, Trait anxiety is associated with hidden state inference during aversive reversal learning. *Nature Communications* **14** (1), 4203 (2023).

Supplementary Materials for

Post-training makes large language models less human-like

This PDF file includes:

Materials and Methods

Data collection

PSYCH-201 was collected through an open research collaboration and crowdsourced effort. We issued a public call on social media inviting researchers to contribute datasets by submitting them to a GitHub repository. Each prompt was designed to include the entire trial-by-trial history of a complete session from a single participant. Human responses within these prompts were marked using the delimiters << and >>. For each dataset, 10% of participants (or up to a maximum of 100) were reserved as a held-out test set. The submitted datasets underwent a lightweight review process to ensure that they did not contain obvious formatting or implementation bugs.

Large language models

We conducted evaluations using three major families on PSYCH-201: Qwen3, a state-of-the-art open-source model family; Llama3.X, a model family with a broad ecosystem; and Olmo3.X, a fully open and reproducible model family. Models were run using the Hugging Face Transformers library in bfloat16 precision. The results reported in the main paper were obtained without applying a prompt template, as we found this setting to yield the best performance, including for models that provide a template. Further results obtained with the default model-specific templates are reported in Fig. S1.

Evaluation metrics

We used average negative log-likelihood over responses as our primary evaluation metric. Evaluation was restricted to the human response spans marked by the delimiters << and >>. For experiments in which a single response consisted of multiple tokens, we first summed the log-likelihoods across all tokens within that response and then averaged these values across responses.

We further quantified post-training misalignment by computing, for each experiment, Cohen's d between the negative log-likelihoods of a base model and its post-trained counterpart, and then averaging the resulting effect sizes across experiments. We report all results on the held-out test set, whose size was sufficient to produce stable estimates.

Data availability

Psych-201 is publicly available on the Huggingface platform at https://huggingface.co/datasets/marcelbinz/Psych-201. The test set is accessible under a CC-BY-ND-4.0 licence through a gated repository at https://huggingface.co/datasets/marcelbinz/Psych-201-test.

Code availability

The code needed to reproduce our results is available at https://github.com/marcelbinz/Psych-201.

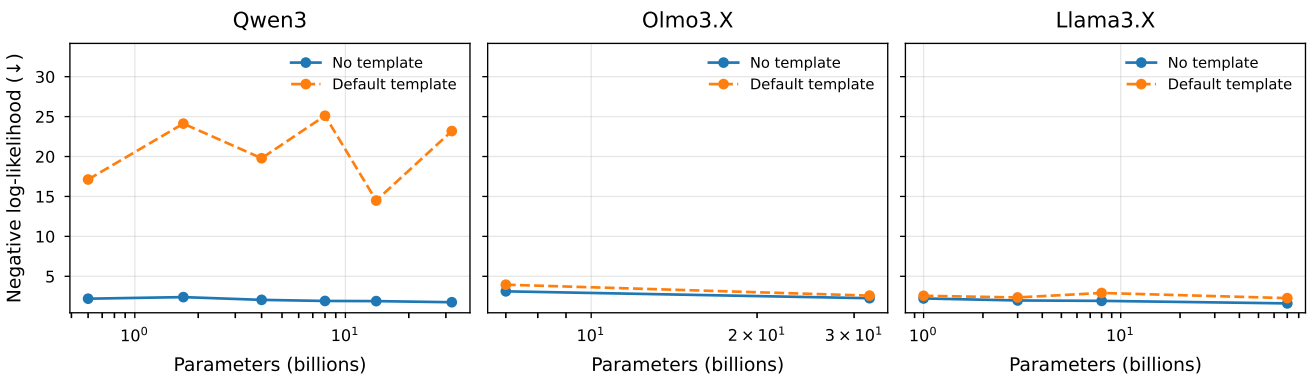

Figure S1: Behavioral alignment for instruction-tuned models prompted with and without chat template. Negative log-likelihood of human responses from PSYCH-201 for models from the Qwen3, Olmo3.X, and Llama3.X families. Models prompted with their default chat template (orange dashed lines) consistently perform worse compared to those without a template (blue solid lines). The biggest gap is observed in the Qwen3 family.

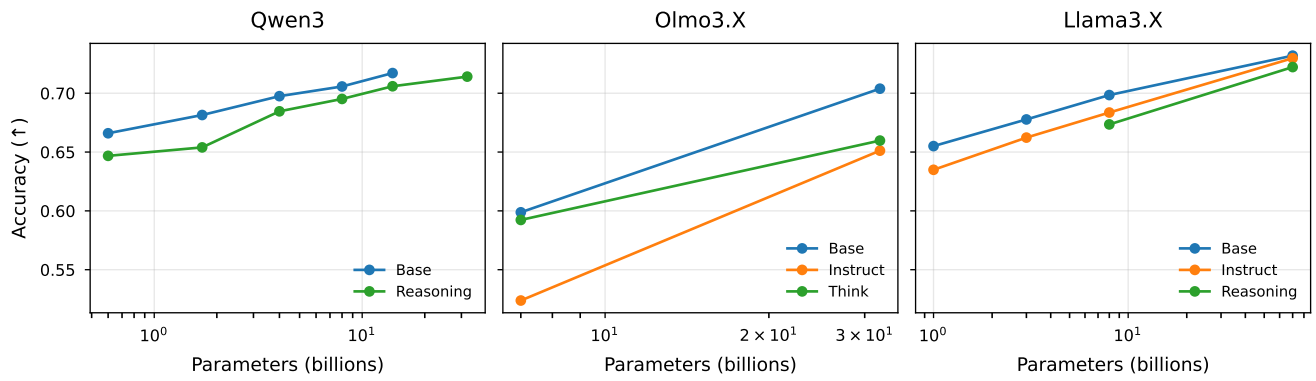

Figure S2: Behavioral alignment for different post-training objectives, measured by accuracy in predicting human responses on a discrete-choice subset of Psych-201 for models from the Qwen3, Olmo3.X, and Llama3.X families. Post-trained models (shown in non-blue) exhibit consistently lower alignment with human responses across all objectives, families, and sizes than their base model counterpart (shown in blue).

Study List (excluding Psych-101)

Study	Domain	Task
`aggarwal2023iag` (*40*)	Reasoning	Inductive argument evaluation
`agrawal2024stress` (*41*)	Intertemporal	Stress and delay discounting
`akata2023repeatedgames` (*42*)	Econ. games	Repeated two-player games
`alsobay2025publicGoodsGame` (*43*)	Econ. games	Public goods game
`anllo2024weird` (*44*)	Reinf. learning	Cross-cultural bandit task
`anvari2024alien` (*45*)	Reinf. learning	Bandit task (alien cover story)
`anvari2024armed_bandit` (*45*)	Reinf. learning	Multi-armed bandit task
`anvari2024observe_bet` (*45*)	Risky choice	Observe-or-bet task
`anvari2024optional_stopping` (*45*)	Risky choice	Optional-stopping (secretary) task
`anvari2024sampling_paradigm` (*45*)	Risky choice	Decisions from experience (sampling)
`awad2018moral` (*46*)	Moral judgment	Moral Machine vehicle dilemmas
`baar2022latent` (*47*)	Econ. games	Social decisions; latent motives
`barnby2022knowing` (*48*)	Econ. games	Social inference; intent attribution
`bavard2018magnitude` (*49*)	Reinf. learning	Instrumental learning; outcome magnitude
`bavard2021range` (*50*)	Reinf. learning	Instrumental learning; range adaptation

Study	Domain	Task
bavard2023functional (*51*)	Reinf. learning	Value normalization in learning
bhatia2024likelihoodratings (*52*)	—	Likelihood/probability ratings
binz2022heuristics (*27*)	Reasoning	Heuristic multi-attribute choice
braendle2023empowerment (*53*)	Reinf. learning	Empowerment-driven exploration
breslav2022shuffle (*54*)	Reinf. learning	Reward learning task
burton2022optimism (*55*)	—	Optimism bias in belief updating
busch2024_navon (*56*)	Cog. control	Navon global–local attention task
busch2024_stroop (*56*)	Cog. control	Stroop interference task
castro_rodrigues2022twostep (*57*)	Reinf. learning	Two-step task
chambon2020feedback (*58*)	Reinf. learning	Learning from free vs. forced choice
cheung2025omissionyesnobias (*34*)	Moral judgment	Moral judgment of omissions
christian2025resolving	Risky choice	Risky choice under uncertainty
ciranka_vandenbos_2024 (*59*)	Risky choice	Social influence on risk-taking
cohen2020causal (*60*)	Reasoning	Causal structure learning
decker2016twostep (*61*)	Reinf. learning	Two-step task across development
demircan2024evaluatingcategory (*62*)	Categorization	Category learning, naturalistic stimuli
demircan2024evaluatingreward (*62*)	Reinf. learning	Reward learning, naturalistic stimuli
dezfouli2019 (*63*)	Reinf. learning	Reward / reversal learning
dubois2022value (*64*)	Reinf. learning	Value-free random exploration
evangelidis2023upscaling (*65*)	—	Preference / choice task
fan2022trait (*66*)	Reinf. learning	Reinforcement learning and traits
feher2020humans (*67*)	Reinf. learning	Two-step task (model-based inference)
franke2024bayesian (*68*)	Psycholinguistics	Bayesian pragmatic inference
frankedegen2016reasoning (*69*)	Psycholinguistics	Pragmatic reasoning; implicature
frey2017dfe (*70*)	Risky choice	Decisions from experience
frey2017lotteries (*70*)	Risky choice	Choices between described lotteries
frey2017mpl (*70*)	Risky choice	Risk elicitation (multiple price list)
gillan2016characterizing (*71*)	Reinf. learning	Model-based control and compulsivity
giron2023developmentalExploration	Reinf. learning	Exploration across the lifespan
gross2023hindsightTransferLearning (*72*)	—	Transfer-learning task
guenther2020LDT 2 (*73*)	Psycholinguistics	Lexical decision task
guenther2020TS (*74*)	Psycholinguistics	Semantic similarity judgments
guenther2022relational (*75*)	Psycholinguistics	Relational similarity judgments
guenther2023ViSpa (*76*)	Psycholinguistics	Visual–semantic similarity judgments

Study	Domain	Task
guenther2023associations_individual	Psycholinguistics	Free word-association generation
guenther2023grammaticality (*77*)	Psycholinguistics	Grammaticality judgments
guenther2024associations_sentences_texts	Psycholinguistics	Word associations to texts
guenther2024comprehension (*78*)	Psycholinguistics	Text comprehension
guenther2024substitutions (*79*)	Psycholinguistics	Word-substitution acceptability
gunadi2021deferral (*80*)	Intertemporal	Choice deferral over time
haines2020intertemporal_choice (*81*)	Intertemporal	Delay-discounting choice
haridi2024memory (*82*)	Memory	Episodic recognition memory
haridi2024memorySemanticContext (*82*)	Memory	Recognition memory; semantic context
haridi2024memoryVisualContext (*82*)	Memory	Recognition memory; visual context
hartley2024twoarmedbandit (*83*)	Reinf. learning	Two-armed bandit task
heffner2022_economicgames (*84*)	Econ. games	Social / economic games
hellmann_unpublished_brightness	Psychophysics	Brightness discrimination
holton2024goalcommitment (*85*)	Reinf. learning	Goal commitment and persistence
hu2023lmpragmatics (*86*)	Psycholinguistics	Pragmatic language understanding
hunter2021increased (*87*)	Econ. games	Social decision-making
hussain2024risk (*88*)	—	Risky choice task
jagadish2023zero (*89*)	Reinf. learning	Zero-shot compositional generalization
jansen2021logic (*90*)	Reasoning	Logical reasoning problems
little2024functionestimation (*91*)	—	Function learning / estimation
marshall_2022_brightness (*92*)	Psychophysics	Brightness discrimination
moutoussis2018pavlovian (*93*)	Reinf. learning	Pavlovian-to-instrumental transfer
nasioulas2024feedback (*94*)	Risky choice	Risky choice with feedback
nussenbaum2020twostep (*95*)	Reinf. learning	Two-step task across development
nussenbaum2023novelty (*96*)	Reinf. learning	Novelty-driven exploration
olschewski2024skewness (*97*)	Risky choice	Experience-based choice; skewness
olschewski2025optimal (*98*)	Risky choice	Information sampling in risky choice
palminteri2017confirmation (*99*)	Reinf. learning	Instrumental learning; confirmation bias
phaneuf-hadd_2025_cogeff (*100*)	Cog. control	Cognitive effort allocation
pike2023catastrophizing (*101*)	Risky choice	Sequential risk task
pirrone_2018_dots (*102*)	Psychophysics	Perceptual decision (dot discrimination)
pirrone_unpublished_food	—	Value-based choice (food)
pirrone_unpublished_lottery	Risky choice	Choices between lotteries
potter2017twostep (*103*)	Reinf. learning	Two-step task

Study	Domain	Task
rausch_unpublished_replication	Reasoning	Reasoning task (replication)
rosenbaum2022valence (*104*)	Reinf. learning	Learning from good vs. bad outcomes
russek2024heuristics (*105*)	Risky choice	Heuristics in sequential choice
rutledge2023happiness (*106*)	Risky choice	Risky choice and momentary happiness
sandbrink2024metacontrol (*107*)	Reinf. learning	Arbitration between control strategies
schiekiera2025metascience (*108*)	Reasoning	Metascientific reasoning
shahar2019twosteptask (*109*)	Reinf. learning	Two-step task
singh2019phishing (*110*)	Reasoning	Phishing-email detection
singh2022representing (*111*)	—	Judgment / representation task
spektor2019contexteffects (*112*)	Reinf. learning	Context effects in repeated choice
spektor2024lossaversion (*113*)	Risky choice	Loss aversion, experience-based choice
sun2025rat (*114*)	Reasoning	Remote Associates Test (insight)
suthaharan2021paranoia (*115*)	Reinf. learning	Reversal learning and paranoia
tesslerfranke2018notunreasonable (*116*)	Psycholinguistics	Interpretation of negated adjectives
thoma2025problearn (*117*)	Reinf. learning	Probabilistic learning task
thoma2025riskychoice (*118*)	Risky choice	Risky choice task
tsvilodub2023xorsome (*119*)	Psycholinguistics	Scalar implicature (or / some)
vandendriessche2022depression (*120*)	Reinf. learning	Reinforcement learning and depression
vantiel2021probabilisticpragmatics (*121*)	Psycholinguistics	Probabilistic pragmatic inference
vantiel2022meaninguse (*122*)	Psycholinguistics	Word meaning and use judgments
witte2024interventionStudy (*123*)	Reinf. learning	Exploration (intervention study)
witte2024safe_exploration (*123*)	Reinf. learning	Safe exploration in bandits
witte_thalmann2024exploration (*124*)	Reinf. learning	Exploration in bandit tasks
xu2021novelty (*125*)	Reinf. learning	Novelty-driven exploration
xu2023augmenting (*126*)	Reasoning	Reasoning with augmentation
ying2023nipe (*127*)	Reasoning	Probabilistic inference task
zhu2024games (*128*)	Econ. games	Economic games
zika2023 (*129*)	Reinf. learning	Learning under uncertainty