What does a language model need to know about a person to predict what they do?

Raj Thaker

WORKING PAPER, NOT PEER REVIEWED · 22 NOVEMBER 2026

Abstract. Language-model "personas" are increasingly used to simulate specific people, usually by conditioning a model on that person's answers to questionnaires. We test what this conditioning buys on 1,190 adults from the Human Connectome Project, scoring every prediction against the real person it was made for and against a supervised model given the same information. Six models from four developers, given a person's measured Big Five, predict that person's other self-report scales nearly as well as the supervised model (mean r = 0.38–0.49 against 0.49), but predict their measured performance on behavioral tasks poorly (0.02–0.09 against 0.17). Asked directly, three Claude models reproduce the true correlations between traits and self-report closely (agreement 0.91–0.97) but not those between traits and behavior (0.44–0.70), with the sign wrong in 31–38% of cases. Adding 16 further self-report scales does not help; adding three short measured cognitive scores raises behavior prediction to 0.26–0.29; supplying the true trait–behavior correlations closes 40–50% of the gap; worked examples help little. We then ask whether brain data can stand in for measured behavior. A text profile decoded from task fMRI lets models predict behavior better than personality does, confirmed on 236 held-out people for four of five models, but nearly all of that signal is also available from how well people performed the tasks in the scanner. Soft-token adapters from brain data into a frozen language model carried a little person-specific information but were beaten by a plain average of the decoded scores. Persona simulations built from self-report reproduce how people describe themselves; to simulate what they do, condition on measured behavior.

1. Introduction

Language models are now routinely asked to stand in for people. They answer surveys as simulated respondents from a demographic group [1], populate simulated towns [2], and serve as "digital twins" of specific individuals, built from interviews or from hundreds of a person's earlier answers [3, 4]. The usual evaluation asks whether the simulated person answers *questions* the way the real one does, and on that test the results are often good: interview-grounded agents reproduce people's General Social Survey answers about 85% as accurately as the people reproduce their own answers two weeks later [3].

Psychology has long distinguished what people say about themselves from what they do. Self-report and behavioral measures of the same construct correlate weakly across many domains [5], partly because behavioral tasks are measured less reliably [6] and partly because they capture different things: a questionnaire records a person's view of their own performance, a task records the performance. A persona built from questionnaire answers might therefore inherit the questionnaire's view of a person: accurate about self-image, much less so about behavior.

We test this directly. For 1,190 adults in the Human Connectome Project (HCP) we have the Big Five, 16 further self-report scales and 11 measured behavioral outcomes (memory, reasoning, attention, emotion recognition and patience for money). We give six language models a person's Big Five and ask for that person's percentile on every outcome. Two controls make the numbers interpretable: a *shuffled* arm that gives each person someone else's profile, and a *supervised ceiling*, a cross-validated ridge regression given exactly the same inputs. The gap between the model and the ceiling, measured separately for self-report and for behavior, says how much of the available information the model actually uses.

Our contributions:

1. **Personas reproduce self-report, not behavior.** All six models come close to the ceiling on self-report and far below it on behavior, with shuffled controls at zero (Section 5.1).
2. **The models' priors are the bottleneck.** Asked for the correlations themselves, models know how traits relate to self-report but not to behavior. Giving them the true table closes 40–50% of the behavior gap (Sections 5.2–5.3).
3. **Measured behavior, not more self-report, is what helps.** Sixteen further questionnaires add nothing; three short cognitive scores roughly quadruple behavior prediction (Section 5.3).
4. **Brain data mostly re-expresses task performance.** A profile decoded from task fMRI helps models predict behavior, confirmed on a sealed set, but in-scanner performance explains nearly all of it. Brain-to-LLM soft-token adapters and a candidate brain-specific decision signal did not hold up (Section 5.4).

We do not claim that the general pattern (personas simulate surveys better than behavior) is new; related evidence exists [3, 4]. What we add is a matched ceiling with shuffled controls on the same people, across models from four developers, a direct probe of the models' beliefs, and a test of which added information closes the gap. We also report several negative results in full.

2. Related work

Language models as simulated people. Argyle et al. condition GPT-3 on demographic backstories and find that it reproduces subgroup attitude patterns [1]. Park et al. build agents with memory and reflection [2] and later ground agents in two-hour interviews with 1,052 people [3]. A pre-registered mega-study on the Twin-2K-500 panel finds that digital twins built from over 500 answers per person correlate only weakly with those people's responses across 164 outcomes (average $r \approx 0.20$), capturing some relative differences but failing at individual prediction [4]. Liu et al. show that models given Big Five responses can role-play answers on nine other psychological scales with realistic inter-scale correlations [7]. Our self-report results agree with this last finding; our behavior results show where it stops.

Ceilings on persona prompting. Hu and Collier measure how much variance persona variables explain in annotation datasets and show that persona prompting is capped by that amount, with a 70B model capturing 81% of what a regression on the same variables captures [8]. We use the same logic, comparing each model with a supervised regression on identical inputs, and split the comparison by outcome type.

Self-report versus behavior. Weak self-report–behavior correlations are well documented [5]. Behavioral tasks often have lower test–retest reliability than questionnaires [6], which caps any predictor of them, including our supervised ceiling.

Brain signals and language models. BrainLLM [9] and MindLLM [10] feed fMRI-derived embeddings into a language model to decode what a person perceived. Both build on learned soft prompts for frozen models [11]. We use the same mechanism for a different target: describing a *person* rather than a *stimulus*, and asking whether that description adds anything to predictions about the person's behavior.

3. Data

HCP Young Adult (S1200, open access) [12]. 1,190 people aged 22–37 with the NEO-FFI Big Five, 16 self-report scales (NIH Toolbox emotion battery and the Pittsburgh Sleep Quality Index) and 11 measured behavioral outcomes: delay discounting for $200 and for $40,000 ("patience for money"), emotion recognition, fluid intelligence, episodic memory, working memory, cognitive flexibility, inhibitory control, verbal memory, spatial orientation and sustained attention. Three further cognitive scores (vocabulary, reading and processing speed) are used only as profile inputs, never as targets.

Task fMRI. HCP's precomputed level-2 contrast maps for seven tasks (28 contrasts), averaged within the Schaefer 400-parcel atlas [13]: 11,200 features for 998 people with all seven tasks. **Resting-state fMRI.** Schaefer-400 functional connectivity (79,800 edges) for 1,017 people.

Basel–Berlin Risk Study [14]. 274 people with the Big Five, 18 self-report scales and 11 behavioral measures, used to replicate the personality results.

UCLA Consortium for Neuropsychiatric Phenomics (OpenNeuro ds000030) [15]. 234 usable people with task-fMRI contrast maps and delay discounting, used for a replication of the brain-specific decision signal that was specified before it was run.

4. Methods

Language-model prediction. Each model is told a person's gender and age band and, depending on the arm, their Big Five percentiles and/or other profile items, and asked for that person's percentile on every target. Models: Claude Opus 5, Claude Sonnet 5 and Claude Haiku 4.5 through the Anthropic API; Llama 3.3 70B [16], Qwen 2.5 32B [17] and Mistral Small 24B served with vLLM [18] on rented GPUs with greedy decoding. Haiku 4.5 ran at temperature 0; Sonnet 5 and Opus 5 do not accept a temperature setting and ran at the API defaults with extended thinking off. Opus 5 was run on 600 people: the first 600, in subject order, with all four resting-state runs (the same 600 used as the test set in Section 5.3).

Controls. The *shuffled* arm gives each person another person's profile with identical wording. The *demographics* arm gives no profile. The *ceiling* is 10-fold cross-validated ridge regression from the same inputs to the same targets.

Metric. Pearson r across people for each target, averaged within target family (self-report or behavior). Intervals and contrasts use a paired bootstrap over people: 2,000 resamples for every analysis except the brain adapters, which use 500 (version 1, single split) and 1,000 (version 1 cross-fitted, and version 2).

Belief probe. Each Claude model is asked to estimate the correlation between each of the 27 targets and each of seven predictors (the Big Five, gender and age band) in this population. We compare the believed and true correlation matrices within each family.

What closes the gap. People are split once into 600 test people (the first 600 with all four resting-state runs) and 590 source people. Arms add, to the Big Five: the 16 self-report scales (behavior targets only); the three cognitive scores; 20 worked examples (source people's profiles with their true outcomes); or a table of true trait–outcome correlations computed on the source people.

Brain-decoded profiles. A ridge decoder trained on development people maps each person's task maps to their Big Five and three cognitive scores. The model is then given these *decoded* percentiles in place of measured ones. To separate brain information from task ability, we compare trained models given in-

scanner task performance (16 accuracy, reaction-time and difficulty variables from the scanned tasks), the task maps, and both.

Soft-token adapters. Task maps are reduced to 256 principal components and mapped by a small network to four soft tokens prepended to a frozen Llama 3.1 8B [16]. The adapter is trained on 22 measures and tested on 8 it never saw (four cognitive, two patience, two self-report). Version 1 uses tokens only; version 2 adds the text profile and learns a residual.

Sealed confirmation. All design choices (the measure list, task maps as the brain input, the profile items, the prompt) were made on 762 development people. The 236 people with complete task maps downloaded afterwards were scored once with three tests written down before the sealed set was scored (in design notes that are not externally timestamped): (1) trained task maps beat personality on behavior; (2) an LLM given a person's own decoded profile beats the same LLM given another person's; (3) the decoded profile beats measured personality.

5. Results

5.1 Personas reproduce self-report, not behavior

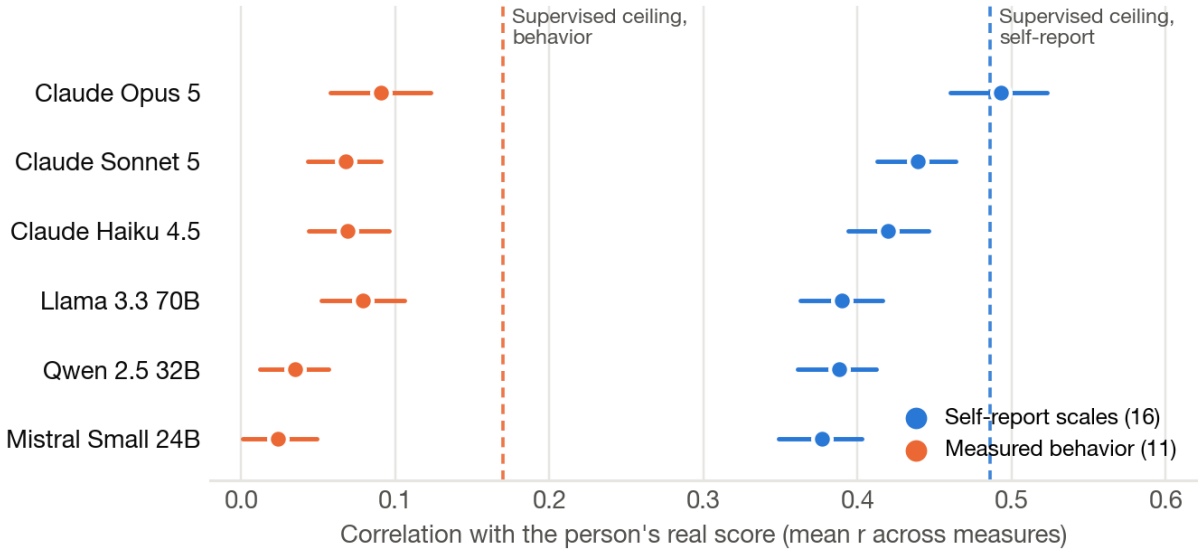


Figure 1. Each model was given a person's measured Big Five and asked for that person's percentile on 16 self-report scales and 11 measured behavioral outcomes. Points are mean Pearson r across people, averaged within family; bars are 95% bootstrap intervals. Dashed lines: ridge regression on the same inputs (n = 1,190). Opus 5 was run on 600 people, whose matched ceiling is 0.52 for self-report and 0.16 for behavior.

Model	Self-report	Behavior	Shuffled, self-report	Shuffled, behavior
Claude Opus 5 (n = 600)	0.49	0.09	0.01	0.02
Claude Sonnet 5	0.44	0.07	0.00	−0.01
Claude Haiku 4.5	0.42	0.07	0.01	−0.01
Llama 3.3 70B	0.39	0.08	0.00	0.01
Qwen 2.5 32B	0.39	0.04	0.01	−0.01
Mistral Small 24B	0.38	0.02	0.02	0.00
Ridge ceiling (n = 1,190)	0.49	0.17		

Against the ceiling for the same people, every model reaches 78–96% of it on self-report and 14–57% on behavior. The shuffled controls are at zero, so the self-report accuracy comes from the person's own profile, not from priors about the sample.

Replication (Basel–Berlin, 274 people). Personality raises self-report prediction above the shuffled control for all five models tested (0.11–0.24, against a ceiling of 0.26). Behavior is near zero for the models *and* for the ceiling (0.02), so this dataset cannot show the gap on its own: personality carries almost no information about its behavioral tasks.

5.2 The models' beliefs about traits and behavior are much less accurate

Model	Agreement with true trait–self-report r	Agreement with true trait–behavior r	Correct sign, behavior
Claude Haiku 4.5	0.91	0.44	62%
Claude Sonnet 5	0.95	0.70	69%
Claude Opus 5	0.97	0.69	64%

The models' believed correlations between traits and self-report scales match the data almost exactly. Their beliefs about behavior match far less well, and the sign is wrong in about a third of cases. Haiku also overstates the size of trait–behavior links (mean believed $|r|$ 0.13 against a true 0.07).

5.3 What information closes the gap

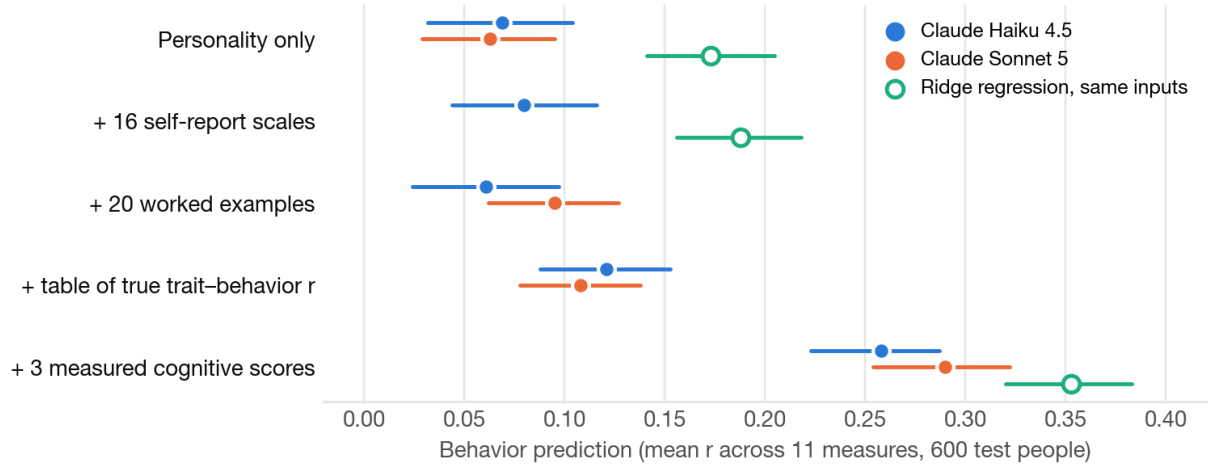


Figure 2. Behavior prediction for 600 test people when information is added to the Big Five. Ridge regression (hollow) is the ceiling for the same inputs; worked examples and the correlation table have no ridge analog. Bars are 95% bootstrap intervals.

Added to personality	Behavior, Haiku 4.5	Behavior, Sonnet 5	Ridge
Nothing	0.07	0.06	0.17
16 self-report scales	0.08	not run	0.19
20 worked examples	0.06	0.10	
Table of true trait–behavior correlations	0.12	0.11	
3 measured cognitive scores	0.26	0.29	0.35

Three findings. More self-report adds almost nothing for the model or the ceiling. The true correlation table helps (Haiku +0.05, Sonnet +0.04), closing 50% and 41% of the gap to the ceiling, so the models' priors are part of the problem. But the largest gain by far comes from measured behavior: three short cognitive scores roughly quadruple behavior prediction for both models. Worked examples help little (Haiku −0.01, Sonnet +0.03).

5.4 Brain data

Trained models. On development people, task maps predict behavior (0.29) but not self-report (0.04); personality shows the reverse (0.16 and 0.50). On the sealed 236 people the pre-specified trained-model test passed: task maps beat personality on behavior (+0.08 [0.03, 0.13]).

Language models given a brain-decoded profile (sealed 236 people). For four of five language models, a person's own decoded profile beat another person's and beat their measured personality (Figure 3). Qwen 2.5 32B failed both tests.

Model	Another person's profile	Personality	Own profile	Own vs another's	Own vs personality
Claude Sonnet 5	0.00	0.09	0.20	+0.20 [0.13, 0.27]	+0.11 [0.05, 0.18]
Claude Haiku 4.5	0.01	0.06	0.17	+0.16 [0.09, 0.23]	+0.11 [0.05, 0.17]
Llama 3.3 70B	−0.01	0.09	0.17	+0.17 [0.10, 0.24]	+0.08 [0.01, 0.14]
Mistral Small 24B	0.03	0.03	0.13	+0.10 [0.03, 0.18]	+0.11 [0.04, 0.17]
Qwen 2.5 32B	0.05	0.03	0.09	+0.04 [−0.02, 0.10]	+0.06 [−0.01, 0.12]

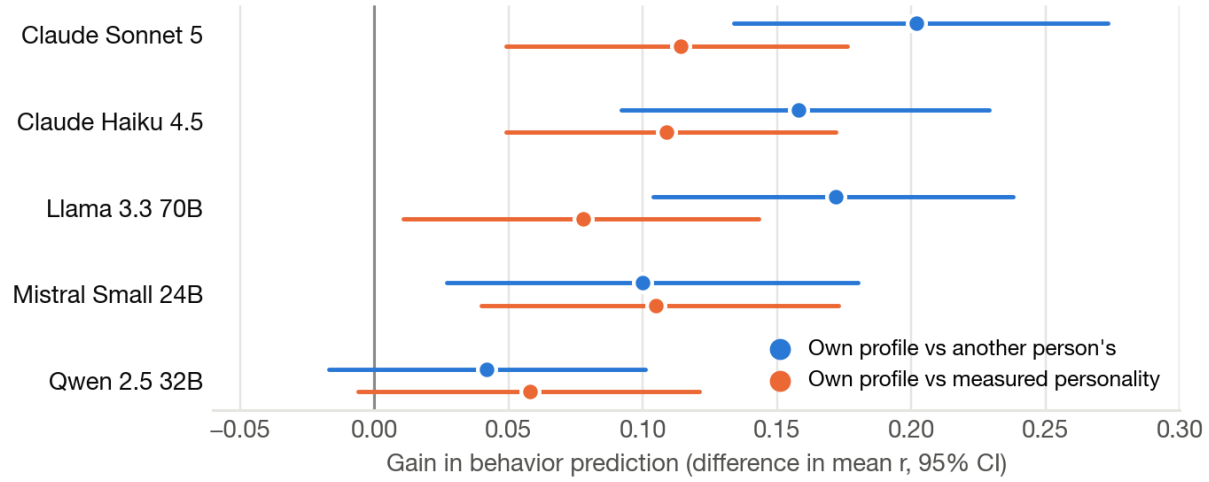


Figure 3. Pre-specified contrasts on 236 sealed people. Points are differences in mean behavior r; bars are 95% paired bootstrap intervals.

Is it the brain, or task ability? The maps are recorded while people perform tasks, and their accuracy and speed on those tasks are themselves strong predictors of the behavioral outcomes. Controlling for this removes most of the brain's contribution (Figure 4). In-scanner performance alone predicts behavior at 0.32; the maps add +0.020 [0.011, 0.029] beyond it, and the maps alone are *worse* than performance alone (−0.054 [−0.076, −0.033]). Resting-state connectivity, which is recorded with no task, predicts behavior at 0.18 on 726 development people (personality: 0.12) and adds +0.017 [0.007, 0.027] beyond in-scanner performance.

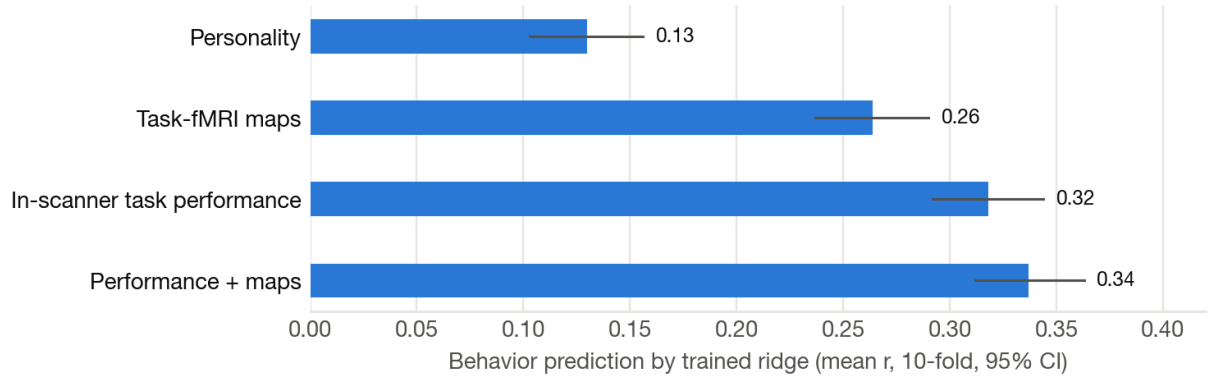


Figure 4. Behavior prediction by trained ridge regression on 748 development people with complete in-scanner data. Task-fMRI maps add little beyond how well people performed the tasks in the scanner.

The same holds for the language models. The decoded profile includes decoded cognitive scores, and the only model with brain-profile predictions on development people (Qwen 2.5 32B) adds +0.008 [0.004, 0.013] beyond in-scanner performance when each outcome uses only the model's own prediction for it (+0.025 [0.015, 0.035] if all of the model's predictions may be used together). A language model can read a brain-decoded profile, but the profile works mainly because task-evoked activity tracks task ability.

A brain-specific decision signal? One outcome looked different. In HCP, task maps predicted patience for $40,000 beyond demographics, personality, all 16 in-scanner performance variables and an overall cognition composite (+0.086 [0.038, 0.133], n = 740); the $200 version pointed the same way but was not reliable (+0.030 [−0.014, 0.073]). We specified a replication on UCLA CNP before running it (the plan is in the project's work log, which is not externally timestamped). It failed (−0.017 [−0.074, 0.037], n = 234). Power was estimated in advance at about 50%, and the datasets differ (no social task, which carried most of the HCP signal; a clinical sample; small-stakes discounting), so the failure is weak evidence against the effect, but the effect cannot be claimed.

Brain adapters. Version 1 (tokens only), on one split of 162 test people, beat shuffled tokens on unseen measures (+0.18 [0.08, 0.28]), with 0.20 on patience. Cross-fitted over all 762 development people it remained person-specific (+0.10 [0.06, 0.14]) but was weak (0.07 on unseen measures, 0.08 on patience) and clearly worse than a plain average of the decoded ability scores (−0.17 [−0.20, −0.13]); these cross-fitted numbers are from the run log, as the analysis prints them without saving a result file. Version 2 (tokens plus text profile, five folds) added no reliable gain over the text profile alone (+0.031 [−0.003, 0.064]), and its gain on patience was not person-specific (own tokens vs another's: +0.007 [−0.046, 0.059]). At this data size, routing brain data through a language model loses information relative to simply averaging what the scan decodes.

6. Discussion

Why the gap exists. Two explanations fit the evidence and are not exclusive. First, the information is not there: personality carries less information about behavior than about self-report, which the ceilings show (0.17 against 0.49). Second, the models do not use the information that is there: they reach 78–96% of the ceiling on self-report but only 14–57% on behavior, and the belief probe shows why. Their beliefs about how traits relate to behavior are poor, and supplying the true relationships recovers about half of what is missing. The likely source is the text these models learn from, in which people describe themselves far more often than their measured performance is reported alongside their personality.

What this means for persona simulation. A persona conditioned on self-report is a model of self-image. That is useful when the goal is to simulate answers to questions, and it is the wrong tool when the goal is to simulate what someone will do. The cheapest fix in our data is not a better prompt or more questionnaire items but a small amount of measured behavior: three short cognitive scores did more than everything else combined.

What brain data adds. Less than it first appears. Brain-decoded profiles helped, survived a sealed test, and worked across model families, which is the result we hoped for. The confound analysis then showed that the same information is available from task performance recorded in the scanner, at no imaging cost. We report this in full because the first-pass result would have looked like a strong positive.

Relevance beyond psychology. Language models increasingly hold representations of the people they talk to. Our results suggest that a model's picture of a person, built from what that person says about themselves, may systematically track self-description rather than behavior. Probing that internal representation directly, rather than through its outputs as we do here, is a natural next step.

7. Limitations

- **Relatives across folds.** HCP includes twins and siblings, and family IDs are restricted-access, so cross-validation folds are not family-aware. This inflates the trained brain results most, especially for heritable outcomes such as delay discounting. The language-model persona results do not involve a trained decoder and are unaffected; the brain-decoded profiles are affected through their decoder.
- **"Behavior" means lab tasks**, mostly cognitive, measured once. Task reliability caps every predictor, including the ceiling [6]. We say nothing about everyday behavior.
- **One main dataset** of healthy young US adults. The personality replication used a second dataset in which behavior was unpredictable for every method; the brain replication used a different, clinical sample.
- **Design changes were made after early looks** (the measure list grew from 4 to 11 behavior outcomes; the brain input changed from resting connectivity to task maps; cognitive scores were added to the profile). They were made on development people and disclosed, and the key language-model brain result was confirmed on a sealed set, but the other analyses are exploratory.
- **The development/sealed split is not distributed.** It was set by download order (the 762 people whose task maps had been downloaded at the time of a snapshot), not by a seeded random draw, and subject lists are not shared. Section 5.4 and the adapter analyses therefore cannot be rerun on exactly the same split.
- **Some arms ran on a subset of models.** Opus 5 ran on 600 people; the information arms ran on two Claude models; the belief probe on three; development-set brain-profile predictions on one open model.
- **Prompting.** One prompt template per arm. Sonnet 5 and Opus 5 ran at default sampling.
- **No head-motion regression** was applied to task maps.
- **Unchecked code.** All code was written and run quickly with AI assistance and has not been independently reviewed.

8. Conclusion

Given a person's personality, current language models predict what that person will say about themselves nearly as well as a supervised model can, and predict what they will do poorly. The models' beliefs about how traits relate to behavior are a large part of the problem, and the most effective remedy in our data was a few minutes of measured behavior. Brain data looked like a second remedy, but most of its value was task performance in disguise. If the goal is to simulate what people do, condition on what they did.

Disclosure, data and code

Code was written with AI assistance (Claude Code). Only derived percentile profiles with no identifiers (gender, age band and percentile ranks) were sent to the Anthropic API; open-weights models ran on

rented GPUs. Every number in this paper comes from the saved aggregate result files in the repository or is printed by its analysis scripts, and figure data are provided alongside the figures. HCP data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657), funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research, and by the McDonnell Center for Systems Neuroscience at Washington University. Under HCP's open-access data use terms, subject-level data are not redistributed. With HCP access the code regenerates every analysis except the development/sealed split, which was set by download order and is not distributed; the repository README lists the analyses that depend on it. This paper has not been peer reviewed.

References

1. 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.
2. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of UIST 2023*.
3. Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. arXiv:2411.10109.
4. Toubia, O., et al. (2025). A mega-study of digital twins reveals strengths, weaknesses and opportunities for further improvement. arXiv:2509.19088 (later versions titled "Digital twins as funhouse mirrors: Five key distortions").
5. Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, 24(4), 267–269.
6. Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *PNAS*, 116, 5472–5477.
7. Liu, Y.-F., Lu, Y.-L., He, D., & Zhang, H. (2026). From five dimensions to many: Large language models as precise and interpretable psychological profilers. *ICLR 2026*. arXiv:2511.03235.
8. Hu, T., & Collier, N. (2024). Quantifying the persona effect in LLM simulations. *Proceedings of ACL 2024*, 10289–10307.
9. Ye, Z., Ai, Q., Liu, Y., de Rijke, M., Zhang, M., Lioma, C., & Ruotsalo, T. (2023). Language generation from brain recordings (later titled "BrainLLM: Generative language decoding from brain recordings"). arXiv:2311.09889.
10. Qiu, W., Huang, Z., Hu, H., Feng, A., Yan, Y., & Ying, R. (2025). MindLLM: A subject-agnostic and versatile model for fMRI-to-text decoding. *Proceedings of ICML 2025*, PMLR 267.
11. Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *Proceedings of EMNLP 2021*, 3045–3059.
12. Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80, 62–79.
13. Schaefer, A., et al. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9), 3095–3114.
14. Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10), e1701381.
15. Poldrack, R. A., et al. (2016). A phenome-wide examination of neural and cognitive function. *Scientific Data*, 3, 160110.
16. Llama Team, Meta (2024). The Llama 3 herd of models. arXiv:2407.21783.
17. Qwen Team (2024). Qwen2.5 technical report. arXiv:2412.15115.
18. Kwon, W., et al. (2023). Efficient memory management for large language model serving with PagedAttention. *Proceedings of SOSP 2023*.